A STUDY OF HUMAN COLLAGEN GENES 
AND THE MOLECULAR GENETICS OF 
OSTEOGENESIS IMPERFECTA

Thesis submitted for the degree of 
Doctor of Philosophy 
at the University of Leicester

by

Nicola Jayne Rose B.Sc. (Hons.) 
Department of Genetics 
University of Leicester

December 1993

CORRECTIONS — ONLY IN THE REFERENCE
COPY OF THIS THESIS
Dedication

This thesis is dedicated to the memory of my grandfather

Francis William Rose,

a man I would very much liked to have known.
A STUDY OF HUMAN COLLAGEN GENES AND THE MOLECULAR GENETICS OF
OSTEOGENESIS IMPERFECTA

Nicola Jayne Rose

The human collagens are a family of related proteins which all possess at least one triple helical domain characterised by a Gly-Xaa-Yaa amino acid repeat motif where Xaa and Yaa are frequently proline and hydroxyproline respectively. This repetitive structure is determined by the characteristic sequence of the genes encoding the component polypeptides. Mutations in the COL1A1 and COL1A2 genes, encoding the proα1(I) and proα2(I) chains of type I collagen have been reported. These mutations can alter the polypeptide chains of type I collagen and interfere with the nature or amount of mature collagen produced by the cell and have been implicated as the cause of the connective tissue disorder osteogenesis imperfecta (brittle bone disease). This disorder is heterogeneous in nature with phenotypes ranging from mild to lethal in the perinatal period.

In this study, single strand conformation polymorphism analysis was coupled with DNA sequencing in order to identify mutations in type I collagen genes which could be responsible for the disorder phenotype in a number of patients. Three novel glycine by serine substitutions were identified, two of which were found to be shared by more than one unrelated individual — a phenomenon rarely observed for osteogenesis imperfecta mutations. The fourth mutation was the first glycine by glutamic acid substitution reported to be the cause of this disorder. All of these substitutions occurred within the triple helical region of the proα2(I) chain. These data add to the knowledge of the relationship between the position and nature of substituting amino acids and the resulting phenotypes. In addition, two novel but neutral sequence variants were identified.

A novel source of collagen mRNAs based upon basal transcription levels was also investigated as was an approach for identifying novel collagenous sequences using degenerate oligonucleotide-primed polymerase chain reaction amplification of DNA.
ACKNOWLEDGEMENTS

Throughout my time in Leicester, a number of people have been involved in the progression of my work. To list everyone would be lengthy but I would like to thank a few people for the major part they played. Katrina Mackay gave me invaluable help in every aspect of my Ph.D. and together with Rachel Carter and past and present members of G18/G26, ensured that I kept my head above water in times of 'crisis' — we even managed to have a few laughs along the way! I thank Dr. Raymond Dalgleish, my supervisor, in whose laboratory I carried out the work for this thesis. Prof. Alec Jeffreys and Drs. Nicola Royle and John Armour for their invaluable guidance and contributions, Maxine Allen for help with the MYR-PCR mapping technique, Matthew Lewis for assistance in collating mutation data, Sarah Laband for the use of her computer on many an occasion and Ian Riddell for expert and prompt photography! I also acknowledge the participation of our collaborators in Europe and the USA as well as the Medical Research Council who funded my three years with a studentship. The EC Concerted Action on "Heritable Connective Tissue Disorders 1990-1992" is acknowledged for supplying the 'LinKit' set of oligonucleotides.

And now the socialites. Three years would not have been bearable without the unforgettable input from people such as Will (little), Will (big hairy), Lee and the rest of the Yeastie boys who showed me the other, lower, side of life; members of the Brain Drain ("minisatellites aren't boring") group; the 1990-91 M.Sc. group and the 'Panto people' who showed me the two (if not more) sides of almost every member of the Genetics Department. Christmas will never be the same again — thankfully! I have shared dwellings with a vast number of people in the space of three years — why so many? Did they know something I didn't? — and they have made my time in Leicester memorable in more ways than one! I would especially like to thank Max, Lee, Will (still little) and David (computer thing) for being complete stars in times of need, stress and general panic! I can always rely on you for choice comments and embarrassing photographs (how much do you want?). And the others — Mat (get ***), Lesley, Mikey D. ("I'm having a crisis"), Mikey H., Tara (Dizzy), Catherine and Eithnie as well as the vast number of '17 St. Peter's Road squatters' who have left their mark on the place. So, thanks for the laughs, the beers and the birthdays that I can't remember let alone forget!

I would also like to thank Chris for opening my eyes and for making the last month of my thesis-writing frenzy a more bearable and enjoyable period of time than I was expecting it to be! And last but not least there are the people who always seemed to know that things would go right in the end — so I thank my grandparents, and for their unquestionable faith in me and tolerance and love over the years I thank Mum, Dad, Steve and Matt.

Thanks.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>I</td>
</tr>
<tr>
<td>CONTENTS</td>
<td>II</td>
</tr>
<tr>
<td>PUBLICATIONS</td>
<td>VII</td>
</tr>
<tr>
<td>ABBREVIATIONS</td>
<td>VIII</td>
</tr>
</tbody>
</table>

## CHAPTER ONE

### INTRODUCTION

1. PREFACE

1.1 CONNECTIVE TISSUES AND THE EXTRACELLULAR MATRIX

1.2 THE FAMILY OF VERTEBRATE COLLAGEN PROTEINS

1.2.1 Fibril-forming collagens

1.2.1.1 Type I collagen

1.2.2 Non-fibril-forming collagens

1.2.3 Fibril Associated Collagens with Interrupted Helices (FACITs)

1.2.4 Invertebrate collagen proteins

1.2.5 The diversity of vertebrate collagens

1.3 FIBRILLAR COLLAGEN BIOSYNTHESIS AND DEGRADATION

1.4 THE COLLAGEN GENES

1.4.1 Distribution in the genome

1.4.2 Fibrillar collagen genes

1.4.3 Non-fibrillar and FACIT collagen genes

1.4.4 Evolution of the collagen gene

1.4.5 Regulation of gene expression

1.4.5.1 Regulation of type I collagen genes

1.4.5.2 Regulation of other collagen genes

1.4.5.3 Basal transcription of collagen genes

1.5 DISORDERS LINKED TO COLLAGEN GENES

1.5.1 Osteogenesis imperfecta

1.5.1.1 The mutations that cause osteogenesis imperfecta

1.5.1.2 Overmodification and locality of mutation

1.5.1.3 Mosaicism

1.5.1.4 Biochemical consequences of OI mutations

1.5.1.5 Social issues and OI

1.5.2 Osteoporosis

1.5.3 Other disorders

1.5.4 The use of mutant animals in the study of osteogenesis imperfecta and other disorders linked to collagen genes

1.6 THE DETECTION OF MUTATIONS

1.6.1 The Polymerase Chain Reaction (PCR)

1.6.2 Mutation detection by conformational changes

1.6.2.1 Single Strand Conformation Polymorphism (SSCP) analysis

1.6.2.2 Denaturing Gradient Gel Electrophoresis (DGGE)

1.6.2.3 Heteroduplex (HD) analysis
1.6.3 Mutation detection by heteroduplex mismatch
1.6.3.1 Ribonuclease (RNAase) cleavage
1.6.3.2 Chemical Cleavage of Mismatch (CCM)
1.6.3.3 Carbodiimide modification (CDI)
1.6.4 Mutation detection by sequencing

1.7 AIMS OF THE STUDY

CHAPTER TWO
MATERIALS AND METHODS

2.1 SOURCES OF MATERIALS
2.2 GENERAL METHODS
2.3 GEL ELECTROPHORESIS
2.4 TRANSFER OF DNA TO NYLON MEMBRANES
2.5 HYBRIDISATION
2.6 RADIO-LABELLING OF MARKER DNA FRAGMENTS
2.7 CLONING OF DNA FRAGMENTS
2.8 ISOLATION OF PLASMID DNA
2.9 PLASMID LIBRARY SCREENING
2.10 NUCLEIC ACID PREPARATION FROM BLOOD
2.11 AMPLIFICATION OF SEQUENCES USING THE POLYMERASE CHAIN REACTION (PCR)
2.12 DNA SEQUENCING
2.13 CULTURE OF FIBROBLAST CELLS
2.14 SSCP ANALYSIS
2.15 MINISATELLITE VARIANT REPEAT PCR
2.16 COMPUTER PROGRAMS

CHAPTER THREE
BACKGROUND TO THE STRATEGY FOR SCREENING FOR MUTATIONS IN PATIENTS WITH OSTEOGENESIS IMPERFECTA

3.1 INTRODUCTION
3.2 DETAILS OF PATIENTS USED IN THE SCREENING STUDY
3.3 THE SSCP STRATEGY
3.3.1 Localisation of the mutation
3.3.2 Identification of the sequence defect
CHAPTER FOUR
IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN TWO INDIVIDUALS WITH TYPE III OSTEOGENESIS IMPEFECTA 57
4.1 INTRODUCTION 57
4.2 DETECTION OF A MUTATION IN PATIENTS PJ AND KO 58
4.3 CHARACTERISATION OF THE MUTATION IN PATIENTS PJ AND KO 58
4.4 DETERMINATION OF THE DISTINCTION BETWEEN THE PATIENT SAMPLES 58
4.5 DISCUSSION 59

CHAPTER FIVE
IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN THREE INDIVIDUALS WITH PERINATAL LETHAL OSTEOGENESIS IMPEFECTA 63
5.1 INTRODUCTION 63
5.2 IDENTIFICATION OF A MUTATION IN PATIENTS DW, PE AND GM02528 63
5.3 CHARACTERISATION OF THE MUTATION IN PATIENTS DW, PE AND 23 64
5.4 DETERMINATION OF THE DISTINCTION BETWEEN THE PATIENT SAMPLES 64
5.5 DISCUSSION 65
5.5.1 Multiple occurrence of mutations 66

CHAPTER SIX
IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN A PATIENT WITH TYPE IV OSTEOGENESIS IMPEFECTA 67
6.1 INTRODUCTION 67
6.2 IDENTIFICATION OF A MUTATION IN PATIENT SC 67
6.3 CHARACTERISATION OF THE MUTATION IN PATIENT SC 67
6.4 DISCUSSION 68

CHAPTER SEVEN
IDENTIFICATION AND CHARACTERISATION OF A NOVEL GLYCINE TO GLUTAMIC ACID MUTATION IN A PATIENT WITH PERINATAL LETHAL OSTEOGENESIS IMPEFECTA 70
7.1 INTRODUCTION 70
7.2 IDENTIFICATION OF A MUTATION IN PATIENT R 70
7.3 CHARACTERISATION OF THE MUTATION IN PATIENT R 70
7.4 DISCUSSION 71
A novel glycine to glutamic acid substitution at position 343 in the α2 chain of type I collagen in an individual with lethal osteogenesis imperfecta.
*Hum. Mol. Genet.* 2: 2175–2177

A Gly859Ser substitution in the triple helical domain of the α2 chain of type I collagen resulting in osteogenesis imperfecta type III in two unrelated individuals.
Rose, N.J., Mackay, K., Byers, P.H., Dalgleish, R. (1994)
*Hum. Mut.* (in press)

Three unrelated individuals with perinatally lethal osteogenesis imperfecta resulting from identical Gly502Ser substitutions in the α2 chain of type I collagen.
Rose, N.J., Mackay, K., De Paepe, A., Steinmann, B., Punnett, H.H., Dalgleish, R.
*Hum. Mol. Genet.* (manuscript submitted)

A Gly238Ser substitution in the α2 chain of type I collagen results in osteogenesis imperfecta type IV.
Rose, N.J., Mackay, K., Byers, P.H., Dalgleish, R.
 manuscipt submitted)
ABBREVIATIONS

Abbreviations throughout this thesis conform to the SI. In addition, amino acids are referred to by the standard single and triple letter codes. Residues in the collagen α-chains are indicated by a code accepted in the field of collagen work i.e. the first glycine of the triple helical domain is classified as number one. Residues preceding this are denoted by negative numbers. Other abbreviations used are listed below.

A adenine
APS ammonium persulphate
bp base pair
BSA bovine serum albumin
cDNA complementary DNA
dNTP 2'-deoxyribonucleotide triphosphate where N is A, C, G or T
ddNTP 2', 3'-dideoxyribonucleotide triphosphate where N is A, C, G or T
C cysteine
C-propeptide carboxyl propeptide
C-terminal carboxyl terminal
CNBr cyanogen bromide
Δ deletion
DEPC diethylpyrocarbonate
dH2O distilled water
DNA deoxyribonucleic acid
DNase deoxyribonuclease
DTT dithiothreitol
EDTA diaminooctanetetra-acetic acid
EtBr ethidium bromide
G guanine
HEPES N-2-hydroxyethylpiperazine-N-2-ethane-sulphonic acid
IAA iso-amyl alcohol
IMS industrial methylated spirits
IPTG isopropyl-β-D-thiogalactopyranoside
kb kilobase pair
mRNA messenger RNA
N-propeptide amino propeptide
N-terminal amino terminal
O.D. optical density
RNA ribonucleic acid
RNase ribonuclease
SDS sodium dodecyl sulphate (sodium lauryl sulphate)
T thymine
TEMED N,N,N',N'-tetramethylethylenediamine
Tris 2-amino-2-(hydroxyethyl)propane-1, 3-diol
U uracil
UV ultra violet
X-gal 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
CHAPTER ONE

INTRODUCTION

PREFACE

Heritable disorders of connective tissue have been studied extensively since the characterisation of the Marfan syndrome and osteogenesis imperfecta. In 1956, the first of four editions of McKusick's *Heritable Disorders of Connective Tissue* was published describing many different disorders. The genetic heterogeneity of several of those disorders has been studied, however, not until 1972 was the first molecular defect underlying a disorder of the connective tissue identified, in Ehlers Danlos syndrome type VI (Krane et al., 1972). Today, the literature regarding molecular defects in many of the connective tissues resulting in medical disorders is vast, mainly as a consequence of the enormous increase in overall knowledge of the disorders and technological advances in molecular biology.

Defects in collagen genes are the main subject of interest of this thesis. However, due to the enormity of the literature and the size constraint imposed upon such a dissertation, it is not possible to provide a comprehensive review of the many themes covered within. Thus for the sake of brevity, reviews of extensive subject matters are referred to instead of the original manuscripts. Since individual works are not cited in the "references" section, the contribution of all authors of such reviewed data is hereby acknowledged.

1.1 CONNECTIVE TISSUES AND THE EXTRACELLULAR MATRIX

The term connective tissue is often used to describe the extracellular matrix (ECM) plus the cells in contact with it such as fibroblasts, macrophages and mast cells. The macromolecules that form the ECM are secreted by local cells. The main components are the collagens, proteoglycans and hyaluronic acid of which the collagens are the most abundant protein. The proteoglycans are hydrophilic and form a highly hydrated gel-like substance in which the collagen fibres are embedded. In addition to these molecules, specialised cells secrete other proteins e.g. elastin which imparts resilience to the matrix; two high molecular weight glycoproteins, fibronectin which is involved in cell adhesion and found in connective tissues as well as blood; and laminin which occurs in basal laminae. Integrins appear to be the major receptors for cell attachment in the ECM and are found in neural tissues. In addition to the commonly found proteins, a plethora of low abundance and less well characterised proteins may also be present, osteonectin in bone for example (Termine et al., 1981).

The relative amounts of matrix components vary enormously as does the manner in which they are organised thus resulting in a great diversity of form and hence function, e.g. the matrix of bone becomes calcified and confers rigidity and strength for bodily support, whereas the matrix in the cornea can assume an almost crystalline, transparent structure. The collagens vary enormously in both form and function and are discussed in the following chapters. The analysis of collagen genes in heritable connective tissue disorders has aided the understanding of the processes underlying triple helical formation and other aspects of collagen biosynthesis.
1.2 THE FAMILY OF VERTEBRATE COLLAGEN PROTEINS

The collagens are a family of structurally related, but genetically and biochemically distinct, proteins each with a characteristic tissue distribution. Collagens are produced by many cell types and constitute approximately one quarter of total body protein in mammals. The proteins are found in essentially all tissues and are involved in maintaining the structural integrity of numerous organs and tissues. They are also involved in early development and organogenesis, cell attachment and platelet aggregation among other functions. A collagen is defined as "a structural protein of the extracellular matrix that contains at least one domain in the characteristic triple-helical conformation" (van der Rest and Bruccoleri, 1993). The collagens have been classified into different types each of which define distinct sets of polypeptide chains which can form homo- or heterotrimeric molecules. To date, 19 types have been identified and the distinct component α-chains are encoded by 32 genes which are dispersed over 12 chromosomes. The different types are denoted by Roman numerals assigned in the order that the collagen was discovered. Table 1.1 shows the distribution of the proteins and genes (section 1.4).

As more collagens were discovered, it was realised that although they might have different structures, function and tissue distributions, similarities existed between them and thus they were divided into groups. Miller (1985) classified the collagens into three groups based on their molecular weight and the supramolecular structure that they assumed in the extracellular matrix, however, this was recognised as an unsatisfactory system and another was proposed by van der Rest and Gairone (1991) which is still in use. They divided the collagens into three groups based mainly on their supramolecular structure. These groups comprised (i) the collagens which formed fibrils in the extracellular matrix, (ii) those that formed other structures and (iii) collagens that did not form fibrils and had interrupted triple helices. These groups will be discussed further. Molecules such as the complement component Clq (Reid et al., 1979) and acetylcholine esterase (Mays and Rosenberry, 1981) have collagenous sequences but they have no known structural role in the ECM and this precludes their classification as collagens.

Each collagen molecule comprises three polypeptide α-chains which may or may not be identical. Each polypeptide forms a left-handed helix; when three α-chains associate every third amino acid of the chain assumes a position in the centre of the superhelix and is shifted by 30° from the preceding central residue from the same polypeptide chain. The overall structure of the triple helical domain is therefore, a right-handed helix. Due to steric hindrance, the third amino acid in each of the α-chains must be glycine since it has the smallest side chain of all the amino acids and is therefore the only one small enough to fit into the centre of the superhelix. The α-chains are thus characterised by a repetitive Gly-Xaa-Yaa amino acid sequence. Any motif other than this triplet amino acid repeat would perturb the triple helix although many collagen triple helical domains do contain irregularities. The α-chains are named according to how many there are for each collagen type e.g. type II collagen has one α-chain — α1(II) — but type IV has six, namely α1(IV) through α6(IV). The triple helix is stabilised by steric repulsion between imino acids in the Xaa and Yaa positions. These imino acids confer rigidity to the molecule and limit rotation about the N-C peptide bond. Some amino acids have their side chains at the surface of the molecule and are therefore thought to be able to form hydrophobic and electrostatic regions that can direct correct self-assembly of the monomer units into fibrils. Approximately one third of all the Xaa and Yaa residues are proline and hydroxyproline respectively (van der Rest and Gairone, 1991) which is unusual since hydroxyproline does not
occur in other proteins and proline is encountered infrequently. In particular, it has been shown that the melting temperature of the triple helix increases with increasing numbers of component hydroxyprolines (Privalov, 1982). The presence of an appropriate number of 4-hydroxyprolines in the Yaa position (approximately 100) has been shown to be crucial for stability since they form extra hydrogen bonds (Kiely et al., 1993). Some lysine and hydroxylysine residues are essential for forming cross-links between adjacent monomer units within a fibril. Such cross-links arise as a result of oxidative deamination of the side chain to form an aldehyde which forms the links. Thus it can be seen that most of the amino acids within the peptide chains of collagen have an important biological function. The triple helix is resistant to proteolytic degradation, however, upon heating the structure undergoes a helix-to-coil transition and is then susceptible to degradative enzymes.

The structure of collagen proteins is such that the amino acids in the Xaa and Yaa positions have their side chains pointing away from the helix and thus may act as substrates for interactions with other extracellular matrix proteins in particular, with other triple helices (van der Rest and Garrone, 1991).

1.2.1 Fibril-forming collagens

Collagens classed together in this group (types I, II, III, V and XI) have the property of being able to form striated fibrils in the extracellular matrix. Initially, it was suggested that the situation arose as a result of the alignment of staggered collagen molecules with a length of approximately 300 nm, producing gaps and regions of overlap (reviewed by Kuhn, 1987). It is now evident that homologous regions within triple helical domains interact in a lateral direction and not in a linear array to produce a quarter-staggered array. The lengths of the overlap and gap zones are 27 nm and 40 nm respectively (Miller, 1985) and the molecules are staggered by approximately 67 nm — a distance termed ‘D’. The fibrils are stabilised by covalent cross-links between different triple helices (Eyre et al., 1984). Distances within the collagen molecule are often quoted using D as a unit length; each molecule being 4.4 D in length. All of the collagens in this group have clearly related genes which appear to be derived from a single common ancestral gene (section 1.4.4). Although these collagens have been shown to form fibrils on their own, they can associate with another collagen type to form a heterotypic fibril (e.g. Mendler et al., 1989; Birk et al., 1988). In order to achieve this, the length of the triple helical domain of the different α-chains must be similar for each fibrillar collagen (section 1.4.2). Different collagens may be found to occur together in the same tissue. For example, type III collagen can be found alongside type I in tissues with the exception of bone and cornea (Fleischmajer et al., 1990). Type V is a relatively minor abundance collagen but is almost ubiquitous in its distribution (Seyer and Kang, 1990) and is found to coexist with types I and III collagen. Types II and XI can both be found in cartilage where type II is the more abundant of the two. The α3(XI) chain is encoded by the same locus as the α1(II) chain but may differ in posttranslational modifications (van der Rest and Garrone, 1991; section 1.2.5).

The details of the biosynthesis of collagens are discussed in section 1.3.
1.2.1 Type I collagen

Type I collagen is the classical fibrillar collagen and has been reviewed by Kühn (1987) and more recently by Kielty et al. (1993). Quantitatively, it is the most abundant collagen being present in almost every organ in the body and representing approximately 90% of total body collagen. Thus it is not surprising that much of our understanding of the structure, synthesis and assembly of collagens is due to studies on type I. It provides the basis of the mechanical strength of structures such as skin, tendon, bone, dentine, cornea and sclera. It is present as large and highly organised fibrils which comprise smaller, identical microfilaments aligned to form a cross-striated pattern seen under electron microscopy. Type I collagen is a semiflexible molecule approximately 300 nm in length with a diameter of about 1.4 nm. It usually exists as a heterotrimer consisting of two α1(I) and one α2(I) chains although a homotrimer of α1(I) chains has been observed in the medium of fibroblast cell cultures (Moro and Smith, 1977; Jimenez et al., 1977; Pihlajaniemi et al., 1984; Blue et al., 1990) and tissues (Rojobind et al., 1979; Uitto, 1979). It has been observed that this homotrimer has an elevated 3-hydroxyproline level and a higher degree of hydroxylation of the lysyl residues on the polypeptide chain than the heterotrimer. Uitto (1979) suggested that although this trimeric form accounts for less than 5% of total collagen in human skin, it may contribute to the tensile strength of various connective tissues and defects in its biochemistry may result in connective tissue disorders. It was observed that the homotrimer formed fibrils at a slower rate in vitro than heterotrimers suggesting that the α2(I) chain is important in driving fibril formation (Kielty et al., 1993). The two type I collagen α-chains are very similar in length but their primary sequences are sufficiently different for them to be separated by ion-exchange chromatography and SDS-polyacrylamide gel electrophoresis.

1.2.2 Non-fibrill-forming collagens

This group of collagens (comprising types IV, VI, VII, VIII and X; reviewed by Mayne and Burgeson, 1987; van der Rest and Garrone, 1991) is very heterogeneous both structurally and functionally; the collagens are involved in the formation of sheets or protein membranes that surround tissues or organs. For example, type IV is the major component of basement membranes; the N-termini join together and the globular C-termini associate with those of other molecules forming a tetramer which entraps large macromolecules. Type VI forms an antiparallel dimer. Two dimers associate along the axis to form a tetramer which interacts in a head to tail arrangement with another tetramer. The resulting filaments can aggregate to form bundles (van der Rest and Garrone, 1991) with a function possibly involved in cell adhesion (Timpl and Engel, 1987). Intramolecular interactions between the triple helices of type VII molecules stabilise the filaments which can be found in anchoring fibres between basement membranes and molecules in the ECM or between ECM molecules themselves. Type VIII, the major component of the Descemet's membrane and other epithelia, forms a hexagonal lattice by association of the termini of the rod-like collagen structures; since the junctions in this lattice appear to be too large if the non-triple helical domain of the type VIII was present alone, it is thought that some other macromolecule may interact with the collagen to produce this structure (van der Rest and Garrone, 1991). The conservation of structure between members of this group is less evident than between fibrillar collagens. Type X is only found in
hypertrophic chondrocytes and is thought to be associated with other proteins (Reginato and Jimenez, 1991) and involved in tissue calcification (Jimenez et al., 1986).

1.2.3 Fibril Associated Collagens with Interrupted Helices (FACITs)

FACTs have been reviewed by Shaw and Olsen (1991) and discussed by Kielty et al. (1993). The collagens in this group are types IX and XII to XVI. All of these collagens possess regions of high homology and have unique structural characteristics not observed in other collagen types. The main difference between the structure of FACTs and other collagen types is the presence of one or more triple helical domains separated by non-triple helical regions. These helical domains often contain cysteine residues which is in contrast to fibrillar collagens where cysteine is excluded from the single triple helical domain (except in type III where cysteine residues are known to participate in cross-linking between individual α-chains of a procollagen molecule (Kielty et al., 1993)). They do not appear to form trimeric molecules which form structures in the extracellular matrix by themselves. Rather they associate with fibrils composed of fibrillar collagens. Type IX collagen is the best characterised of this group and is the only one discussed in detail here. A functional role for the FACTs is proposed based on the studies of this collagen. Type IX collagen consists of three triple helical domains (COL1, COL2 and COL3) and three interspersed non-collagenous domains (NC1 to NC3) in the α2 and α3 chains; the α1(IX) chain possesses an additional NC4 domain. The NC3 domain of the α2(IX) chain is five residues longer than the NC3 domain on α1(IX) and α3(IX) and contains an attachment site for a proteoglycan side chain (Huber et al., 1986; McCormick et al., 1987). It is thought that these extra amino acids cause the type IX molecule to kink (Irwin and Mayne, 1986). Vaughan et al. (1988) demonstrated that the short arm of the collagen (the COL3 domain) is at an angle to the rest of the molecule. The COL1 and COL2 domains associate with the surface of a fibrillar collagen structure such that the short arm projects from the surface. This could determine the interaction of collagen fibrils with other fibrils or components of the extracellular matrix. Type IX has been found in association with type II in cartilage and it may be that the kinked part of the molecule projects into the gap region of the type II fibrils (Bruckner et al., 1988). This collagen has also been seen to exist in another, shortened, form which arises from the use of a second transcription start site. The structural difference between the two forms could influence the nature of the interaction between itself and yet another collagen. The shortened molecule has been detected in developing chick embryonic cornea as well as mouse and foetal human tissues (Muragaki et al., 1990).

Type XII has been seen to be associated with type I in dense connective tissue structures such as tendon and ligament. It has been suggested that it interacts through the triple helical regions leaving its largest non-collagenous domain (NC1) free to project into the ECM and to interact with other molecules (Kielty et al., 1993).

1.2.4 Invertebrate collagen proteins

Several collagen genes have been isolated from a few invertebrate species and some are presented here. Aspects of the subject have been reviewed by Har-El and Tanzer (1993). The nematode Caenorhabditis elegans produces several different cuticles throughout its life cycle and a large family of collagen proteins (50–100 separate
types) has been found in tissues from the organism. Several collagen genes and proteins have also been isolated from *Drosophila melanogaster* and genes from two species of sea urchin (reviewed by Chu and Prockop, 1993) have been discovered as well as two species of deep sea worm (Gaill et al., 1992). Tauzes et al. (1993) report the existence of an mRNA in a marine annelid which appears to be related to human type IX collagen. Most of these collagens appear to be analogous to human non-fibrillar collagens with one known exception (Exposito et al., 1992) which is analogous to the human α2 chain of type I collagen. The details and interactions of these proteins and their corresponding genes will not be discussed further.

1.2.5 The diversity of vertebrate collagens

The vertebrate collagens are by no means a homogeneous set of proteins; in fact there are several areas of diversity within the family. At the DNA level a number of distinct collagen genes exist, these are discussed in section 1.4. Since most collagen genes contain numerous introns (except for the genes encoding types VIII and X, the short chain collagens) it is not surprising that alternatively spliced exons in the transcripts of some collagens have been observed. Various patterns of alternative splicing exist and vary from single exon exclusion to differential use of exons through alternative promoters. This provides yet another degree of diversity of the proteins. Type XIII undergoes the most extensive alternative splicing and is the only collagen that has been shown to undergo such splicing events in the triple helical domain. The six possible transcripts have distinct tissue distributions (Juvonen et al., 1992). Many of the collagens gene transcripts possess several polyadenylation sites which give rise to more than one mRNA species (Aho et al., 1983; Myers et al., 1983). *In vitro* ratios of the different size mRNAs can be altered by the addition of transforming growth factor-β although the *in vivo* significance of this is unknown (Penttinen et al., 1988). Two different forms of type IX collagen exist with widely different amino terminal globular domains (Svoboda et al., 1988) which result from two different transcription start sites within the COL9A1 gene (Nishimura et al., 1989). The structural differences between the two forms and the tissue distribution could alter the interaction between the type IX collagen molecule and other fibrils and thus the protein function. The region of the COL9A1 gene encoding the cysteine-rich region of the NC4 domain of the α1(IX) chain undergoes alternative splicing through the use of two different promoters. Such splicing differences result in the production of a normal α1(IX) chain and an alternative α1(IX) chain which differs through the possession of an extended NC4 domain; both forms are present in foetal and juvenile chondrocytes but the alternatively spliced form is more abundant in foetal cells (Bruckner et al., 1988). The function of the alternative form has not yet been elucidated. At the level of chain assembly there is another degree of diversity; it has been noted that collagen chains can be a part of different trimeric molecules. This varying chain stoichiometry is evident in type 1 collagen in which the molecular species can be a homotrimer consisting of only α1(I) chains or a heterotrimer consisting of α1(I) and α2(I) chains. Type V collagen has many forms utilising its three different α-chains in various combinations. Type IV collagen has five different α-chains and thus the number of possible different trimeric molecules is numerous although many of these forms may be of minor or specialised importance. The α2 chain of type I collagen has also been shown to participate in the formation of a type XI collagen. This
phenomenon of sharing chains between collagen types has also been suggested between types V and XI in bovine bone (Niyibizi and Eyre, 1989); leading to the classification of some collagen molecules as isomers.

Diversity among the proteins can also be observed at the supramolecular level. This is indicated by evidence showing that collagen molecules often interact with other extracellular components in order to achieve the structural properties required by the cell. Covalent cross-links between types I and III collagen can occur (Henkel and Glanville, 1982; Eyre et al., 1984). It may be that pN collagen III (a common species of type III collagen which has retained the N-propeptide) may regulate the diameter of type I collagen fibrils (Fleischmajer et al., 1981) perhaps due to such cross-links. Romanic et al. (1991) demonstrated that pN collagen II may inhibit the rate of assembly and thus incorporation with fibres of type I collagen in vitro. Type V may have a similar function; structurally it is very similar to the fibrillar collagens. This phenomenon can also be seen in the FACITs (section 1.2.3).

1.3 FIBRILLAR COLLAGEN BIOSYNTHESIS AND DEGRADATION

In order to discuss the biosynthesis of fibrillar collagens it is necessary to introduce the basic structure of the component α-chains. Each chain consists of multiple Gly-Xaa-Yaa repeats in which Xaa is often proline and Yaa is often hydroxyproline. This repetitive primary structure is reflected at the gene level, the composition of which is discussed in full in section 1.4.

The biosynthesis of fibrillar collagen genes is discussed with respect to type I collagen. Fibrillar collagens are synthesised initially as a procollagen species with a large globular polypeptide extension at each end and are the soluble precursors of the different collagen types. The process is complex and multi-stage, beginning with the translation of the N-terminal signal peptide which is essential for efficient and selective targeting of nascent polypeptide chains to the endoplasmic reticulum (ER), and the subsequent translocation into the lumen of the ER. Following the translation of the total mRNA encoding the pre-proα collagen chains on the ribosomes, a non-collagen-specific signal peptidase cleaves the signal peptide to yield a proa-chain. These polypeptides have a molecular weight approximately 50% greater than that of the mature α-chain due to the presence of large globular polypeptide extensions at both the amino- and carboxyl- termini. The amino (N-) terminal propeptide comprises a cysteine-rich domain, a short triple helical domain and a short globular region. The function of the cysregion has not yet been elucidated but its removal may be necessary for fibril formation (Kühn, 1987). Another non-helical region, the amino telopeptide, separates the N-terminus from the main triple helical region of the collagen molecule. The N-terminal propeptide of the α2(I) chain is shorter than the equivalent region of the α1(I) chain since it lacks the very N-terminal globular domain. The carboxyl (C-) propeptide is a globular domain containing cysteines and is involved in inter-chain association.

The underlying processes involved in helix formation and secretion are only partially understood. The process is initiated by the association of the C-terminal propeptides of three α-chains. The process of chain selection is thought to involve the attachment of the polypeptides at a common site at the endoplasmic reticulum (Kirk et al., 1987). Chain association — probably initiated by non-covalent interactions between C-terminal
propeptides due to the association of aromatic and hydrophobic residues within the C-terminus (Brass et al., 1992) which are stabilised by disulphide bonds — and alignment, takes place as the chains reach their final length and occurs first at the C-terminal ends. This process may be facilitated by the organisation and position of the ribosomes on the ER. Binding through the C-terminal propeptides ensures correct registration of the α-chains. Incorrect registration leads to regions of mismatched triple helix. Non-fibrillar collagens possess naturally occurring short interruptions of the Gly-Xaa-Yaa motif or a discontinuous triple helical domain, but these do not disrupt the correct folding of the collagen molecules. It has been proposed that the formation of the helix proceeds via nucleation and propagation events. There is a five-fold repeat of the Gly-Pro-Hyp motif which occurs at the very C-terminus of the collagenous domain in type I collagen (Kielty et al., 1993). Such sequences are known to form the most stable helical structure and thus suggest a highly stable nucleation site. Brass et al. (1991) proposed a three-way aromatic zipper effect whereby the aromatic side groups which represent six of the twelve completely conserved amino acids at the C-terminus, act as teeth in a zipper-like mechanism. This nucleation event initiates helix winding which then proceeds by propagation towards the N-terminal ends of the chains and results in the chains being staggered by one amino acid; an arrangement known to be necessary for the glycines to be packed into the centre of the helix. The resulting helical structure is stabilised by hydrogen bonds and electrostatic interactions as well as a large number of 4-hydroxyprolines (approximately one for every ten prolines) in the Yaa position. Modifications to various residues of the transcribed sequences occur as co-translational processes while the polypeptide is extending on the ribosomes but the reactions are continued until triple helix formation of the α-chains occurs. The main modifications are the hydroxylation of prolyl and lysyl residues and the glycosylation of hydroxylysyl residues. Hydroxylation involves three hydroxylases, prolyl 4-hydroxylase, prolyl 3-hydroxylase and lysyl hydroxylase which react with the sequences Xaa-Pro-Gly, Pro-4Hyp-Gly and Xaa-Lys-Gly to yield 4-hydroxy-L-proline, 3-hydroxy-L-proline and 5-hydroxy-L-lysine respectively. The process of glycosylation is effected by hydroxylysyl galactosyl transferase which modifies hydroxylysyl residues by the addition of galactose, and galactosylhydroxylysyl glucosyl transferase which modifies galactosyl hydroxylysine residues by the addition of glucose. All the enzymes involved only recognise nascent α-chains as substrates for the reaction. Once a triple helix is formed, modifications of that region cease. The processes involved in both modes of modification have been reviewed (Kielty et al., 1993). Molecules that possess insufficient numbers of modified residues accumulate in the lumen of the ER. Research into the processes underlying triple helical formation has been carried out by studying naturally occurring mutant collagens isolated from individuals with inherited collagen disorders (section 1.5). The principle of nucleated growth has been reviewed by Prockop (1990) and Engel and Prockop (1991). The secretion of procollagen molecules into the extracellular space is also poorly understood. The rate is apparently dependent upon triple helical conformation. If this is prevented, the non-helical proα-chains accumulate in the cisternae of the rough ER and are secreted at a slower rate. Upon secretion of procollagen molecules and prior to fibril formation, the N- and C-terminal propeptides are cleaved by procollagen N- and C-proteinase respectively. Both of these enzymes are dependent upon the associated conformation of the chains and the availability of Ca^{2+} ions. C-proteinase cleaves the type I (hetero- and homotrimers), type II and type III collagens; however, a different enzyme is involved in the cleavage of the N-propeptide of type III collagen (Kühn, 1987;
Dombrowski and Prockop, 1988) than of types I and II collagen. Kadler et al. (1987) demonstrated that the removal of the amino-terminus with N-proteinase resulted in so-called pC-collagen molecules that did not assemble into fibrils or aggregate into a mass. The only structures of this nature seen to form occurred when trace amounts of mature collagen or pN-collagen molecules (molecules that have retained the N-terminal region) were present in the sample. The subsequent removal of the C-terminus from purified pC-collagen with C-proteinase resulted in the formation of fibrils, a process seen to be reversible with the lowering and subsequent raising of temperature. These experiments demonstrated that the removal of the C-terminus is the critical step for fibril formation.

Upon removal of the N- and C-terminal propeptides, the resulting molecule consists of a large triple helical region flanked by short, non-helical telopeptide domains. The spontaneous aggregation of processed collagen molecules into fibrils with a D-periodicity is followed almost immediately by the formation of covalent cross-links within and between collagen molecules. These are formed by the oxidative deamination of specific lysyl and hydroxylysyl residues to aldehydes. These aldehydes react with adjacent amino groups and form intermolecular cross-links which afford tensile strength and mechanical stability to collagen fibrils. The processing of other collagens differs, due to their function in the ECM and the supramolecular structure they adopt. The FACITs are probably not processed at all (Shaw and Olsen, 1991). The collagen may then interact with components of the ECM, e.g., the possible inclusion of hydroxyapatite crystals into the gaps of type I fibrils which may afford additional strength to the structure (Prockop, 1979).

The triple helix is highly resistant to enzyme degradation due mainly to extensive cross linking in and between molecules in the fibrils. The binding of carbohydrate molecules to the fibres and the interaction with glycoproteins in the ECM heightens this resistance. Therefore, the cellular turnover of collagen is slow. When degradation of the protein is required, e.g., during tissue (re)modeling, a pre-degradation event occurs. Connective tissue cells secrete a group of matrix metalloproteinases (MMPs) comprising a variety of different molecules. Included in the group are interstitial and neutrophil collagenases. These cleave the triple helix towards the C-terminus; the helical structure rapidly disintegrates and the unwound chains become viable substrates for non-specific proteases (reviewed by Murphy and Reynolds, 1993).

1.4 THE COLLAGEN GENES

Recent reviews discuss the collagen genes in detail (Vuorio and de Crombrugghe, 1990; Chu and Prockop, 1993). Collagen genes are indicated by the prefix COL followed by an Arabic number to denote the type of collagen and then by a Roman character and a second Arabic numeral to indicate the α-chain, e.g., COL1A2 refers to the α2 chain of type I collagen and COL4A5 refers to the α5 chain of type IV collagen.

1.4.1 Distribution in the genome

The chromosomal assignment of 27 of the human collagen genes has now been defined. Despite the collagens being part of a protein family, their genes are not clustered in the human genome as the α- and β-globin genes are, for example, on chromosomes 16 and 11 respectively. The genes encoding the various α-chains of the
human collagens are, in the main, widely scattered throughout the genome e.g. the loci COL1A1 and COL1A2 are non-syntenic, residing on chromosomes 17 and 7 respectively. However, five exceptions exist. The genes for the α1 and α2 chains of type IV collagen are closely associated the genes being on opposite strands of DNA on chromosome 13 in a head to head arrangement; the first exons of the two genes are separated by no more than 130bp (Boyd et al., 1988). Type VI genes are closely associated on chromosome 21 (Weül et al., 1988), and α1(III) and α2(V) both map to 2q24.3-q31 (Emanuel et al., 1985); they lie in a tail to tail arrangement separated by 35kb. The loci COL4A3 and COL4A4 also appear to be closely associated on chromosome 2 (Mariyama et al., 1992) and there appear to be two genes in close association on chromosome 6, namely COL9A1 and a recently discovered gene similar to the COL9A1 gene identified as the D6S228E locus for the α1(Y) collagen chain (Yoshioka et al., 1992). The recently discovered α6 chain of type IV collagen (Zhou et al., 1993) appears to be closely located to the COL5A4 gene and the two may share a common promoter. Details of all the known genes are given in table 1.1.

1.4.2 Fibrillar collagen genes

Fibrillar collagen genes have a well defined structure. Sequence comparisons have been made between genomic and cDNA clones (Boyd et al., 1980, Ohkubo et al., 1980) which identified intron/exon boundaries and demonstrated exon size similarities between different collagen types. Although all the fibrillar collagen genes exhibit a high degree of similarity, the gene structure for fibrillar collagens is discussed with respect to types I, II and III. The basic structure consists of 52 exons of which exons 7 through 48 code entirely for triple helical sequences as do parts of exons 6 and 49. In the proα(I) gene, the 54bp exons 33 and 34 are fused. One of the features of the genes is that the triple helical domain is encoded by 44 exons which are strikingly similar between the different collagens, thus the fused exons in the proα(I) gene are referred to as exon 33/34 (Mouson et al., 1982; Ramirez et al., 1985) so that exons from different collagen genes can still be aligned and directly compared. Similarly, in the proα(III) gene, exons 4 and 5 are fused and are referred to as exon 4/5. In type II, two exons encode the N-terminal domain one of which appears to be an alternatively spliced exon (Ryan et al., 1990).

The 5' untranslated exon sequences of fibrillar genes are relatively short. The N-propeptide and the signal peptide are encoded by six exons which do not display regions of high conservation (Vuorio and de Crombrugghe, 1990). The short triple helical domain does not show size constraints as seen in the main helical region and its function has not been elucidated. It may be that it is a non-functional remnant of evolution. The C-propeptide is encoded by exons 49–52. The first two exons can be variable in size resulting in size differences between different collagens whereas the last two exons are highly conserved in both size and sequence in and between species (Bernard et al., 1983a, b).

Sequence analysis has revealed that the exons appear to be multiples of 54bp or combinations of 45bp and 54bp. In all cases each exon starts with a complete codon for glycine, ends with a complete codon for a Y amino acid and encodes a discrete number of Gly-Xaa-Yaa repeat units. The terminal of the triple helix are encoded by so-called joining exons which also encode telopeptide and propeptide sequences. Differences in the number of the Gly-Xaa-Yaa repeats in these joining exons account for the minor variation in the length of the triple helical
<table>
<thead>
<tr>
<th>COLLAGEN TYPE</th>
<th>CONSTITUENT CHAINS</th>
<th>GENE LOCUS</th>
<th>CHROMOSOMAL LOCATION</th>
<th>MOLECULAR SPECIES</th>
<th>TISSUE DISTRIBUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>α1(I)</td>
<td>COL1A1</td>
<td>17q21.3-q22</td>
<td>[α1(I)]2α2(I)</td>
<td>Skin, bone, tendon, dentin</td>
</tr>
<tr>
<td></td>
<td>α2(I)</td>
<td>COL1A2</td>
<td>7q21.3-q22</td>
<td>[α1(I)]3</td>
<td>Dentin, skin</td>
</tr>
<tr>
<td>II</td>
<td>α1(II)</td>
<td>COL2A1</td>
<td>12q13-q14</td>
<td>[α1(II)]3</td>
<td>Hyaline cartilage, vitreous humour</td>
</tr>
<tr>
<td>III</td>
<td>α1(III)</td>
<td>COL3A1</td>
<td>2q24.3-q31</td>
<td>[α1(III)]4</td>
<td>Skin, vessels</td>
</tr>
<tr>
<td>IV</td>
<td>α1(IV) α2(IV) α3(IV) α5(IV) COL4A1 COL4A2 COL4A4 COL4A5 COL4A6</td>
<td>13q34 13q34 2q35-37 2q35-37 Xq22 X</td>
<td>[α1(IV)]α2(IV) and minor forms</td>
<td>Basement and glomerular membranes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Foetal oesophagus, adult kidney</td>
</tr>
<tr>
<td>V</td>
<td>α1(V) α2(V) α3(V) COL5A1 COL5A2 COL5A3</td>
<td>9q34.2-q34.3 2q24.3-q31</td>
<td>[α1(V)]α2(V)α3(V)</td>
<td>[α1(V)]3</td>
<td>Vessels, skin, intervertebral disc</td>
</tr>
<tr>
<td>VI</td>
<td>α1(VI) α2(VI) α3(VI) COL6A1 COL6A2 COL6A3</td>
<td>21q22.3 21q22.3 2q37</td>
<td>[α1(VI)]α2(VI)α3(VI)</td>
<td>[α1(VI)]3</td>
<td>Anchoring fibrils in dermal-epidermal junctions</td>
</tr>
<tr>
<td>VII</td>
<td>α1(VII) COL7A1</td>
<td>3p21</td>
<td>[α1(VII)]3</td>
<td>[α1(VII)]3</td>
<td>Descemet’s membrane, endothelium, mesenchyme</td>
</tr>
<tr>
<td>VIII</td>
<td>α1(VIII) α2(VIII) COL8A1 COL8A2</td>
<td>3q12-13.1 1p32.3-p34.3</td>
<td>unknown</td>
<td>unknown</td>
<td>Dense connective tissue, tendon, ligament, bone</td>
</tr>
<tr>
<td>IX</td>
<td>α1(IX) α2(IX) α3(IX) COL9A1 COL9A2 COL9A3</td>
<td>6q12-q14 1 unknown</td>
<td>[α1(IX)]α2(IX)α3(IX)</td>
<td>[α1(IX)]3</td>
<td>Hyaline cartilage, vitreous humour</td>
</tr>
<tr>
<td>X</td>
<td>α1(X) COL10A1</td>
<td>6q21-q22</td>
<td>[α1(X)]3</td>
<td>[α1(X)]3</td>
<td>Hypertrophic cartilage</td>
</tr>
<tr>
<td>XI</td>
<td>α1(XI) α2(XI) α3(XI) COL11A1 COL11A2 COL12A1</td>
<td>1p21 6p21.2 12q13-q14</td>
<td>[α1(XI)]α2(XI)α3(XI)</td>
<td>[α1(XI)]3</td>
<td>Cartilage, vitreous humour</td>
</tr>
<tr>
<td>XII</td>
<td>α1(XII) COL12A1</td>
<td>6</td>
<td>[α1(XII)]3</td>
<td>[α1(XII)]3</td>
<td>Embryonic tendon and skin Peridental ligament</td>
</tr>
<tr>
<td>XIII</td>
<td>α1(XIII) COL13A1</td>
<td>10q22</td>
<td>unknown</td>
<td>unknown</td>
<td>Endothelial cells</td>
</tr>
<tr>
<td>XIV</td>
<td>α1(XIV) COL14A1</td>
<td>unknown</td>
<td>[α1(XIV)]3</td>
<td>[α1(XIV)]3</td>
<td>Foetal skin and tendon</td>
</tr>
<tr>
<td>XV</td>
<td>α1(XV) COL15A1</td>
<td>9q21-q22</td>
<td>unknown</td>
<td>unknown</td>
<td>Fibroblasts, keratinocytes</td>
</tr>
<tr>
<td>XVI</td>
<td>α1(XVI) COL16A1</td>
<td>1p34-p35</td>
<td>unknown</td>
<td>unknown</td>
<td>Skin hemidesmosomes</td>
</tr>
<tr>
<td>XVII</td>
<td>α1(XVII) COL17A1</td>
<td>1q42.3</td>
<td>unknown</td>
<td>unknown</td>
<td>Liver, kidney, placenta</td>
</tr>
<tr>
<td>XVIII</td>
<td>α1(XVIII) COL18A1</td>
<td>unknown</td>
<td>unknown</td>
<td>unknown</td>
<td>Rhabdomyosarcoma cells</td>
</tr>
</tbody>
</table>

Table 1.1 The distribution of collagen genes and proteins.

Information regarding collagen genes and proteins identified to date is illustrated. The data have been collated from Kivirikko (1993), van der Rest and Garrone (1991) and Zhou et al. (1993). Genes which have been shown to be in close association with each other are indicated by matching symbols.
domains between α1(I), α2(I) and α1(II) (1014 residues) and α1(III) (1029 residues). The exons are separated by introns of variable size.

1.4.3 Non-fibrillar collagen and FACIT genes

The exon structure and gene organisation of the non-fibrillar collagens differ from the fibrillar collagen genes although in general, the basic 9bp unit encoding the Gly-Xaa-Yaa motif is maintained. A comparison of genomic sequences of different non-fibrillar collagens shows little homology between their coding domains. The degree of divergence between genes within this group is probably related to the diverse structure and function of these collagens compared to that of fibrillar collagens. Analysis of clones of the genes encoding the proc1(IX) and proc2(IX) collagen chains revealed that the number and size of exons is distinct from that in fibrillar collagen genes (Lozano et al., 1985). Type IV is the major non-fibrillar collagen and the structure of the α1(IV) gene is strikingly different from fibrillar genes in that just two of the exons are 54bp in length and only three are 45bp in length (Soininen et al., 1989). In addition, only the first 19 exons of the triple helical domain begin with complete codons for glycine, thereafter most begin with a split codon for glycine; the first G of the codon residing in the preceding exon; split glycine codons are not found in fibrillar collagen genes. However, the number of exons in the triple helical domain is consistent with that in fibrillar genes. Only partial information is currently available regarding the gene structure of other non-fibrillar genes. The α2(VI) chain consists of six exons for the large C-terminal globular domain; the triple helix is encoded by approximately 20 exons that are multiples of 9bp but the majority are 63bp in length and thus the 54bp exon does not appear to be as common. The paucity of exons 54bp in length is evident in some of the genes for the FACITs; chick α1(IX) has only three, chick α1(XII) has one, chick α1(X) has none and human α1(XXIII) has just two (Chu and Prockop, 1993). In each of these genes there is no predominant size of exon.

1.4.4 Evolution of the collagen gene

Fibrillar collagen gene exon structure appears to be conserved throughout evolution. With the exception of the fused exons, the four genes for type I, II and III collagen α-chains have the same corresponding exon sizes. Exon size is also conserved between man, rodent and chick which suggests a common ancestor in the gene structure. This similarity extends to the N- and C-propeptides. Since the majority of exons are 54bp in length, it has been speculated that this basic exon unit may represent the evolutionary origin of present day exon structure of the collagens, in particular the fibrillar collagens (Yamada et al., 1980; Wozney et al., 1981). Through a series of tandem duplications, transpositions to adjacent non-collagenous coding sequences and unequal crossing overs, this primordial sequence may have become the multixenon gene family present today. Exons of 108bp and 162bp may have resulted from loss of intervening introns whereas exons of 45bp and 99bp could have arisen by recombination events between two 54bp exons or deletion of Gly-Xaa-Yaa motifs, although it is not clear why these two sizes would have been generated in preference to others. Exposito et al. (1993) suggested that a possible ancestral collagen gene may have consisted of two types of unit, namely the aforementioned 54bp unit and a 45bp-54bp
tandem unit and that these two motifs were duplicated in a similar fashion as discussed, to yield different collagen genes.

The possibility of selective pressure maintaining the gene structure exists possibly because of rigorous length requirements for the collagen molecule in the extracellular matrix. In general, it is less likely that recombination events take place between exons than introns, however, due to the repetitive nature of collagen exons, such recombination events may occur more often. It has been suggested that the high intron/exon size ratio is necessary to reduce the incidence of unequal crossing over between repetitive sequences (reviewed by Sandell and Boyd, 1990). Rapidly diverging intron sequence would be a feature of the primitive collagen gene due to the unequal crossing over and the loss of sequence is not improbable. This could explain the fusion exons seen in the pro-α1(I) and pro-α1(III) genes. Intron loss has also been observed between species i.e. man, mouse and chick (Upholt and Sandell, 1986). The chromosomal dispersion of the genes may also serve to reduce the likelihood of unequal recombination during meiosis within fibrillar collagen genomic sequences where differences in the length of the resulting protein product would be deleterious or lethal to the organism (Ramirez et al., 1985; Solomon et al., 1985). The exact nature of the ancestral exon unit is still unknown and many investigators have proposed sequences based on the frequency of occurrence of different motifs in various genes (reviewed by Sandell and Boyd, 1990).

Non-fibrillar collagen genes are certainly ancient; proteins have been found in Drosophila and nematodes as well as man (section 1.2.4) and thus suggest an age of at least 700 million years (Hostilka and Tryggvason, 1987). Data suggest that fibrillar collagen genes are not a recent divergence from the primordial gene (reviewed by Sandell and Boyd, 1990). The pattern appears to have been established before the evolutionary divergence of invertebrates and vertebrates since sea urchins have been shown to possess a collagen gene with a very similar structure (Fields, 1988; D’Alessio et al., 1989; Exposito et al., 1993). The presence of split glycine codons in non-fibrillar collagen genes adds further evidence for a distant evolutionary divergence of the two classes of genes. From work carried out on type IV collagen, Buttice et al. (1990) suggested that such interrupted codons may have evolved from introns by the inactivation of splice site sequences such that the intronic sequence becomes incorporated into exon sequence. Their data also supported the existence of a primordial 54bp exon unit.

It therefore appears that the 54bp exon motif is highly conserved and may reflect an invariant primordial sequence that has been selected for, and has survived throughout evolution. This would suggest that such a highly structured gene may be intolerant of any mutations; a more actively evolving gene may be expected to have a higher mutational load threshold. The variety of mutations in these genes that result in a number of medical disorders (section 1.5) strengthens the debate in favour of an ancient gene.

1.4.5 Regulation of gene expression

Regulation of expression of collagen genes has been reviewed (Sandell and Boyd, 1990; Ramirez and Di Liberto, 1990). The vertebrate collagens have a diverse set of functions being involved in structural systems, developmental programs (cell adhesion, cell movement) and physiological processes (wound healing, hemostasis). The number of types of collagens with different structural functions and defined tissue distribution reflects such a
plethora of functions. To complement this diversity, a number of regulatory sequences for specific expression patterns and thus collagen synthesis must exist.

The transcription of collagen genes is complex. The region of the gene thought to be most important in the regulation of expression is the 5' end. Efficient control is dependent on a series of interactions between DNA cis-acting regulatory elements and available trans-acting nuclear factors that have positive or negative effects on gene expression. In general, for collagens with a wide bodily distribution, such as type I, gene expression must be regulated in order to co-ordinate the ubiquitous production of the two component α-chains and control the amount of protein made. On the other hand, for collagens with a restricted tissue distribution such as type II, the regulation must act to restrict its expression in most tissues.

Promoter and first intron sequences have been investigated in several collagen genes and certain motifs have been found which may not be sufficient for transcription but may be able to bind trans-acting nuclear factors. CpG islands have been proposed as regulatory sequences (reviewed by Sandell and Boyd, 1990). Methylation in collagen genes may be complex but some studies have shown that the CpG dinucleotide may act as a substrate for methylation and that such modifications can influence transcription rates, seemingly by inducing negative control. Enhancer elements have been identified in the first intron of several collagen genes.

The regulation of collagen synthesis can also occur at the post translational modification stage, secretion, processing and assembly steps (section 1.3).

1.4.5.1 Regulation of type I collagen genes

The most common species of type I collagen is the heterotrimeric molecule. The transcription of the two non-syntenic genes for the two α-chains must be tightly controlled in order to maintain correct levels of each chain so that the correct stoichiometry is attained. The regulation of type I collagen genes has been reviewed (Slack et al., 1993). Both type I collagen genes have characteristic TATA and CCAAT sequences within 100bp of the 5' end signifying the presence of promoter elements. However, the promoter sequences differ between the two genes. Only one common binding factor (the CCAAT box binding factor, CBF) interacts with both the α1(I) and α2(I) promoter sequences. Studies using the murine α1(I) and human α1(I) promoters (Slack et al., 1993) have identified regions of the gene capable of directing transcription of a reporter gene. Much work has been carried out using chimeric constructs in NIH-3T3 fibroblasts and Xenopus oocytes to show that sequences 5' to the α1(I) promoter may contain both positive and negative transcription regulatory sequences that act in a tissue-specific manner. The cis-acting elements in the α2(I) promoter have been reviewed (Ramirez and Di Liberto, 1990). Studies, similar to those above, identified regions enhancing transcription in NIH-3T3 fibroblasts although the manner of interaction of such sites and trans-acting nuclear factors remains unknown.

One of the first indications that an intronic sequence could have a functional effect in transcription regulation came from studies of the murine Mov-13 strain. This strain had the Moloney murine leukaemia virus (M-MuLV) inserted into the first intron of the pro-α1(I) collagen gene which, although did not disrupt transcriptional regions or splice sites, removed a DNase-I-hypersensitive site 5' to the start of transcription (Harbers
et al., 1984). This site is associated with transcription of the normal collagen gene. The Mov-13 strain has a perinatal lethal phenotype resulting from a homozygous mutation (section 1.5.4).

There have been several suggestions that the first intron in type I collagen genes may contain cis-acting elements which modulate the rate of transcription. The controversy over the nature and effect of such sequences has arisen due to their unusual characteristics. They appear to be orientation-dependent, a characteristic not observed with other cis-acting elements, and are sensitive to position (Sherwood and Bornstein, 1990). Rippe et al. (1989) found sequences within the first intron of COL1A1 that could exert both positive (enhancer) and negative (silencer) effects on transcription. One appeared to have quite a strong effect — a 3-4-fold increase in promoter activity was witnessed in the pro-α1(I) gene with an enhancer (Bornstein et al., 1988). Much work has been carried out on the murine α2(I) collagen gene to identify sites involved in the transcription process and its regulation (reviewed by Ramirez and Di Liberto, 1990). Although mouse genes have been used to investigate the α1(I) first intron, rodent and human α2(I) first introns do not display large sequence identity, thus it is possible that the human α2(I) gene may contain regulatory elements completely different to those found in the mouse or that the regulatory sequences may be located in regions other than the 5' flanking sequence and first intron.

Relative activities of the promoter and intronic elements have been ill-defined since most of the investigative work on these sequences has used fibroblast cell lines which do not offer an in vivo system. Although type I collagen is found in most tissues in the body, there are cells which do not produce the protein. To achieve such tissue-specificity, there must be a tight control on the interaction between DNA-binding proteins which may be expressed or active in certain cells only, and the cis-acting elements within the gene. In the retrovirally mutated Mov-13 mouse strain discussed above, Hartung et al. (1986) found that the insertion inhibits the start of transcription in fibroblasts and other mesenchymal tissues but not in odontoblasts (Kramochkov et al., 1989) or osteoblasts (Schwarz et al., 1990). It is possible that these cell types control transcription through different mechanisms utilising different cis-acting elements although no evidence for this exists.

Developmental control of gene expression has also been observed in the chick (Merlino et al., 1983). A co-ordinate enhancement of type I, II and III collagen synthesis occurred between 5 and 15 days of development correlating with the development of major collagen-containing tissues. Results from a series of experiments by Bornstein and McKay (1988), suggest that multiple interactions occur between a set of trans-acting factors — which binds to the promoter and 5' flanking sequence of the collagen sequence, and a second set of factors that binds to intronic sequences.

1.4.5.2 Regulation of other collagen genes

Of the other collagens, the regulation of type IV collagen genes has been investigated the most and is the best understood (reviewed by Ramirez and Di Liberto, 1990) although the complete process is still unknown. This is the only process discussed here. The main species of type IV collagen is one consisting of two α1 and one α2 chains. The genes involved, COL4A1 and COL4A2, lie in a head to head arrangement on chromosome 13 in both mouse and human and are co-ordinately regulated during embryogenesis when the deposition of type IV collagen is critical for cell migration and differentiation (Timpl and Dziadek, 1986). The approximately 150bp region that
separates the two genes is highly conserved between mouse and humans and contains a putative binding site for SP1 within a region of dyad symmetry representing a bi-directional promoter for these two genes. In the mouse, high transcription levels are maintained only if enhancer elements in the first intron of COL4A1 are present (Burbele et al., 1988) and in humans, a negative cis-acting element is present in the third intron of COL4A2 (Pöschl et al., 1988). Thus it appears that the regulation of transcription of type IV collagen genes may be dependent on the interaction of both positive and negative cis-acting sequences. It is possible that the COL4A4 and COL4A5 genes and the type VI collagen genes also share regulatory sequences. It is not only the genes for different α-chains of the same collagen that may be co-regulated, Yoshioka et al. (1992) identified the D6S228E locus for the α1(Y) collagen (table 1.1) which may be co-regulated with COL9A1 if it becomes evident that the two proteins interact at a supramolecular level.

1.4.5.3 Basal transcription of collagen genes

Also known as “illegitimate” or “ectopic” transcription, basal transcription is the phenomenon whereby tissue-specific genes are transcribed at a very low level in non-specific cells. Using PCR amplification, Chelly et al. (1988) detected dystrophin transcripts in tissues not known to possess the protein as well as cultured non-muscle cells such as lymphoblastoid, hepatoma and fibroblast cell lines. They suggested that in these non-tissue-specific cells, the transcript was present at approximately one copy per 500–1000 cells which is physiologically insignificant. Basal transcription has now been recognised for a number of different gene transcripts including the cystic fibrosis trans-membrane conductance regulator gene (CFTR; Fonknechten et al., 1992), α1(II) collagen (Chan and Cole, 1991), and α5(IV) collagen (Kaebelehan et al., 1992). The phenomenon has implications in the study of inherited disease (reviewed by Kaplan et al., 1992) since the ability to amplify transcripts from any tissue source could potentially circumvent the need for affected tissue for analysis and would provide a non-invasive diagnostic approach for the screening of known mutations occurring in gene coding regions. The latter is exemplified by the characterisation of a novel point mutation in the factor VIII mRNA causing haemophilia A (Berg et al., 1990). This subject will be expanded upon in Chapter Ten.

1.5 DISORDERS LINKED TO COLLAGEN GENES

It is not surprising that many disorders have been found to result from defects in collagen genes due to the elaborate structure of these genes and the precise synthesis and supramolecular structures of collagen or enzymes necessary for correct processing of collagen. Both a deficiency and an excess of collagen formation in organs can result in medical problems. For example, a deficiency might delay wound/fracture healing and an excess might cause fibrotic conditions. Mutations have been identified in seven of the 32 collagen genes, namely COL1A1, COL1A2, COL2A1, COL3A1, COL4A5, COL7A1 and COL10A1. However, this does not imply that deleterious mutations in the other collagen genes do not exist; many mutated genes could be the underlying cause of disease in which various tissue fragility is a phenotype. Of the disorders found to be linked to collagen genes, osteogenesis imperfecta and the Ehlers-Danlos syndromes are the best studied. Osteogenesis imperfecta is a subject
of analysis in this thesis and is described in detail below. The Ehlers-Danlos syndromes and some of the other disorders are described in brief in section 1.5.3.

Since it has become evident that many of the disorders caused by defects in connective tissues are highly heterogeneous in nature, an international phenotypic nosology has been formulated (Beighton et al., 1988) which defines, among other diseases, those resulting from defects in collagen genes. A corresponding molecular nosology has also been produced (Beighton et al., 1992) which represents a consensus on the underlying molecular defects in some collagen disorders.

1.5.1 Osteogenesis imperfecta

Osteogenesis imperfecta (OI; brittle bone disease) was originally described as either a tarda or congenita form in the nineteenth century but is now recognised as being a disorder which displays enormous variation of phenotype. The disease is characterised, in the main, by brittle bones but other tissues rich in type I collagen can also be affected to varying degrees depending on the type of OI. Such tissues include skin, ligaments, tendons, fasciae, sclerae, teeth and the middle and inner ear. Due to the heterogeneity of the disorder, a classification system to aid diagnosis was introduced by Sillence et al. (1979) who divided the disorder into four main groups. Each of these groups is highly heterogeneous and some further sub-classifications have been introduced (Sillence et al., 1984, 1986). However, the four main types remain unchanged. Type II OI is the most severe form and is usually lethal in utero or shortly after birth, although some infants with type II/III have survived for months or even a number of years with intensive protection and medical supervision (e.g. the patient described by Wang et al., 1992). Type III is the most severe of the non-lethal forms while types IV and I are moderate and mild forms respectively. The clinical features of each class of OI are summarised in table 1.2. In all its forms OI probably affects between 1 in 5000-10000 individuals regardless of racial or ethnic groups however, there may be some forms that are more common in isolated regions possibly due to the accumulation of a rare recessive allele. This was an argument invoked to explain the increased occurrence of type III OI in the black South African population (Beighton and Versfeld, 1985; Viljoen and Beighton, 1987). In most instances though, the phenotype is a consequence of dominant mutations in either of the two genes encoding type I collagen, COL1A1 and COL1A2, and consistent linkage of dominantly inherited OI to these two genes has been shown (Sykes et al., 1990). Only one example of a recessively inherited OI mutation has been documented (Denk et al., 1983; Dickson et al., 1984; Nicholls et al., 1984; Pihlajaniemi et al., 1984). Exceptions to linkage to type I collagen include the homozygous individual documented by Aitchison et al. (1988) born to consanguineous parents. It was suggested that the genes controlling the posttranslational modification enzymes might be candidates for harbouring the mutation causing the phenotype since the mutation was recessive. Another example of OI found not to be linked to either of the type I collagen genes is that reported by Wallis et al. (1993) which may offer an alternative explanation for the high OI III occurrence in the black South African population (see above). This elevated frequency suggests another mechanism for the recurrent phenotype in families other than parental gonadal mosaicism (section 1.5.1.3). Once again, it can be postulated that genes for posttranslational modification enzymes or those enzymes involved in collagen processing may be involved.
<table>
<thead>
<tr>
<th>OI TYPE</th>
<th>CLINICAL FEATURES</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>
| 'mild'  | Normal or near normal stature  
          Variable number of fractures prior to puberty, usually starting after learning to walk  
          Blue sclerae lightening gradually prior to adulthood  
          Hearing loss in 50%  
          Dentinogenesis imperfecta rare  
          Sub-groups:  
          IA — no dental abnormalities  
          IB — dentinogenesis imperfecta present  |
| II      |                  |
| 'lethal' | Lethal in the perinatal period  
          Minimal calvarial mineralisation; broad, beaded ribs; beaked nose; compressed femurs; bowed tibias; marked long bone deformity; flexed and abducted hips; small thoracic cavity; platyspondyly  
          Dark sclerae  
          Prematurity and low birth weight  
          More than 60% die within 24 hours, ~80% die within 1 month  
          Sub-groups:  
          IIA — short, broad, “crumpled” long bones; bowed tibias; continuously beaded ribs  
          IIB — short, broad, “crumpled” femora; angulation of tibias; normal or incompletely beaded ribs  
          IIC — long, thin, inadequately modelled, rectangular long bones with multiple fractures; thin beaded ribs  |
| III     |                  |
| 'severe' | Very short stature  
          Progressively deforming bones which may or may not be present at birth; undermineralised calvarium; thin ribs  
          Head may be large but mental impairment not common  
          Sclerae variable in hue at birth and in infancy  
          Long term survival not uncommon  
          Hearing loss common  
          Dentinogenesis imperfecta common  |
| IV      |                  |
| 'moderate' | Variable short stature — generally short for the family  
           Mild to moderate bone deformity; variable fracture rate; scoliosis in some  
           Sclerae may be blue at birth but become normal  
           Hearing loss in some  
           Dentinogenesis imperfecta common  
           Sub-groups:  
           IVA — no dental abnormality  
           IVB — dentinogenesis imperfecta present  |

Table 1.2 The clinical features of osteogenesis imperfecta

The clinical features observed for all types of osteogenesis imperfecta are presented. This is the Sillence classification (Sillence et al., 1979) with later modifications (Sillence et al., 1984, 1986; reviewed by Byers, 1993) although it is generally recognised that there are regions of phenotypic overlap between classes.
In some cases the patient presents with dentinogenesis imperfecta — malformation of the teeth — and this is the basis of some of the sub-groups found in OI types.

1.5.1.1 The mutations that cause osteogenesis imperfecta

There are several classes of mutation that have been recognised as being responsible for an OI phenotype. These are discussed below. The various consequences of such mutations are discussed in section 1.5.1.4.

(a) Point mutations

This type of mutation is the most common one that results in an OI phenotype. Mutations are denoted by the chain in which the change occurs followed by the amino acid present in the normal individual and its position along the chain, and finally by the substituting residue. For example, α1(I)G (or Gly) 499 C (or Cys) represents the substitution of the glycine in position 499 of the α1(I) chain by cysteine. The occurrence of a single base change in either a consensus donor or acceptor splice site results in an exon skipping event (sec (b)). If the mutation alters an important site within the carboxyl-terminal propeptide domain, the ability of the chains to associate with each other, prior to helix formation, may be affected. This is illustrated by an α1(I)G1017C substitution that results in a mild OI phenotype (Cohn et al., 1988; Labhard et al., 1988) perhaps because the chain is excluded from mature collagen molecules. It is likely that the mutation occurred in a non-conserved region of the carboxyl-terminal domain. In contrast, an arginine for leucine substitution at amino acid 1210 in the conserved region of the carboxyl-terminal domain results in a lethal phenotype probably because the secretion of the abnormal procollagen is affected (Chessler et al., 1993).

The type of point mutation which is best characterised is that which occurs in glycine codons of the triple helical domain. In general, point mutations which affect X and Y position amino acids in the triple helix have no phenotypic effect, although the substitution of a highly conserved arginine at position 618 in the α2(I) chain was seen to result in a variant of the Marfan syndrome (Phillips et al., 1992). However, when the mutation occurs in either the first or second position of a glycine codon, the presence of the different amino acids within the α-chain can have dramatic, often lethal, effects. This is a consequence of the bulkiest amino acid interfering with helix formation. Helix propagation proceeds as normal from the C-terminus towards the N-terminus. However, when the substituted amino acid is encountered in place of a glycine, helix formation is delayed momentarily in order to incorporate the larger residue into the centre of the triple helix. As a consequence of this delay, the nascent portions of the α-chains are open to overmodification by the processing enzymes. The biochemical consequences of this are outlined in section 1.5.1.4.

From the accumulation of data concerning point mutations at glycine codons, it has been possible to demonstrate the existence of a relationship between genotype and phenotype for certain amino acid substitutions. As a very general rule, substitutions for glycine residues result in the most severe OI phenotype when they occur towards the carboxyl-terminal end of the molecule and a milder phenotype when they occur towards the amino-terminal end (Starman et al., 1989). However, the position of substitution is not the definitive determining
factor and the nature of the substituting amino acid is important as may be local domain effects (Sarkar et al., 1992a; Sheffield et al., 1993). The effects of different amino acid substitutions along the chain can be seen by investigating different substituting amino acids at the same position on the α-chain. Several such instances have been reported on the α1(I) chain (see figure 1.1); G85R and G85V both result in a mild phenotype, G382C and G382S in a moderate form and G565V and G565S cause lethal 01. However, other examples cause, in one case very, different phenotypes as illustrated in the following list: G415C (moderate/severe)/G415S (lethal/severe), G541D (lethal)/G541S (severe), G883S (mild)/G883D (lethal), G1009S (severe)/G1009V (lethal). No such identical position substitutions have been documented for the α2(I) chain. The possible existence of domains or co-operative blocks has been documented for other proteins by Privalov (1982) as well as specifically for collagens, suggesting that the denaturation of many proteins with internal homology cannot be considered to be a two-state transition from the native to denatured species. In collagens, these domains are probably delimited according to the residues in the Xaa and Yaa positions (see also Chapter Six) since these are the variable factors in the polypeptide sequence and the nature of such residues may influence the strength of local electrostatic interactions. With regard to collagen substitutions, it was noted that a substitution in a region of the α-chain predicted to result in a certain phenotype on the basis of surrounding substitution/phenotype data, may result in a different phenotype in the individual. It has been suggested that certain regions of the collagen molecule may be susceptible to denaturation even though the temperature required to denature the whole molecule has not been attained (Kadler et al., 1988) resulting in a portion of the helix unwinding at a different temperature than the molecule as a whole. Once again, the neighbouring X and Y position residues are likely to be the main factor in influencing such premature unwinding. Such regions have been termed micro-unfolding units and if, as has been suggested, these units occur in the same locality as a co-operative block, substitution by an amino acid at a specific glycine residue within the collagen molecule may not be identical in its phenotypic effect as the substitution of another glycine by the same amino acid (Pack et al., 1989; Westerhausen et al., 1990; Wenstrup et al., 1991; Marini et al., 1993a) and different amino acids substituted for glycine will have different phenotypic effects. This theory is substantiated by the different effects of serine, and to a lesser extent cysteine, substitutions for glycine and serves to explain the occurrence of lethal phenotypes N-terminal to non-lethal ones. The chain in which the mutation occurs may also affect the resulting phenotype due to the principle of procollagen suicide (section 1.5.1.4).

Data regarding glycine substitutions are listed in tables 1.3 (COL1A1) and 1.4 (COL1A2). In addition, to illustrate the genotype:phenotype link, substitution maps are illustrated in figures 1.1 (COL1A1) and 1.2 (COL1A2). Of the 20 naturally occurring amino acids, mutations in a codon for glycine will only result in the substitution of a limited number of residues. The codon GGN which encodes a glycine residue, can be mutated in any of the three nucleotide positions but only those occurring in the first two positions will have any effect on the amino acid. Base changes in the first position can give rise to arginine, cysteine, serine, tryptophan or a stop codon. Those occurring in the second position result in the substitution of the glycine residue by alanine, aspartic acid, glutamic acid or valine. To date, the substitution of a glycine by tryptophan has not been documented. The failure to identify a glycine by tryptophan substitution is not unexpected since the amino acid is encoded by a single codon (TGG) which would require a G to T transition to take place in the first position of the GGG codon for glycine. The
<table>
<thead>
<tr>
<th>MUTATION</th>
<th>PHENOTYPE</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>G19C</td>
<td>osteoporosis</td>
<td>Kuivanen et al. (1991) FASEB J. 6: 2052-2060</td>
</tr>
<tr>
<td>G65C</td>
<td>mild</td>
<td>Mottes et al. (1993) Fifth Int. Conf. on OI abst. 82</td>
</tr>
<tr>
<td>G85L</td>
<td>lethal</td>
<td>Deak et al. (1991) J. Biol. Chem. 266: 21827-21832</td>
</tr>
<tr>
<td>G85V</td>
<td>(same patient)</td>
<td>Mottes et al. (1993) Fifth Int. Conf. on OI abst. 60</td>
</tr>
<tr>
<td>G94C</td>
<td>lethal</td>
<td>Dyne et al. (1993) Fifth Int. Conf. on OI abst. 116</td>
</tr>
<tr>
<td>G97D</td>
<td>severe</td>
<td>Lightfoot et al. (1992) J. Biol. Chem. 267: 25521-25528</td>
</tr>
<tr>
<td>G308C</td>
<td>mild</td>
<td>Byers et al. (1990) Trends Genet. 6: 293-303</td>
</tr>
<tr>
<td>G325C</td>
<td>lethal</td>
<td>Byers et al. (1990) Trends Genet. 6: 293-303</td>
</tr>
<tr>
<td>G525V</td>
<td></td>
<td>Patterson et al. (1989) J. Biol. Chem. 264: 10083-10087</td>
</tr>
<tr>
<td>G598R</td>
<td></td>
<td>Byers (1993) (see &quot;references&quot; section)</td>
</tr>
<tr>
<td>G479C</td>
<td>severe</td>
<td>Mottes et al. (1993) Hum. Mut. 2: 196-204</td>
</tr>
<tr>
<td>G539S</td>
<td>severe</td>
<td>Gomez-Lira et al. (1993) Fifth Int. Conf. on OI abst. 120</td>
</tr>
<tr>
<td>G559D</td>
<td>lethal</td>
<td>Byers (1990) Trends Genet. 6: 293-303</td>
</tr>
<tr>
<td>G579S</td>
<td>moderate</td>
<td>Gomez-Lira et al. (1993) Fifth Int. Conf. on OI abst. 120</td>
</tr>
<tr>
<td>G598S</td>
<td>lethal</td>
<td>Westerhausen et al. (1990) J. Biol. Chem. 265: 13995-14000</td>
</tr>
<tr>
<td>G637V</td>
<td>severe</td>
<td>Tsucheyoshi et al. (1991) J. Biol. Chem. 266: 15608-15613</td>
</tr>
<tr>
<td>G691C</td>
<td></td>
<td>Kaufmann et al. (1991) FASEB J. 6: 2052-2060</td>
</tr>
<tr>
<td>G832S</td>
<td>moderate</td>
<td>Marini et al. (1989) J. Biol. Chem. 264: 11893-11900</td>
</tr>
<tr>
<td>G844S</td>
<td>severe</td>
<td>Pack et al. (1989) J. Biol. Chem. 264: 10694-10699</td>
</tr>
<tr>
<td>G847R</td>
<td></td>
<td>Wallis et al. (1990) J. Biol. Chem. 265: 18628-18633</td>
</tr>
<tr>
<td>G913S</td>
<td></td>
<td>Cohn et al. (1990) Matrix 10: 236 (abstract)</td>
</tr>
<tr>
<td>G928A</td>
<td></td>
<td>Lamande et al. (1989) J. Biol. Chem. 264: 15809-15812</td>
</tr>
</tbody>
</table>

...cont.
Table 1.3 Glycine substitutions in the COL1A1 gene

Glycine substitutions characterised in the COL1A1 gene are listed with the resulting phenotype. Unless the mutations are discussed further in the text, references are given in shortened form and are not included in the REFERENCES section. Mutations are indicated by the glycine residue number followed by the substituting residue. Amino acids are indicated by the standard single letter code. Where no phenotype is given it is the same as the preceding one.

<table>
<thead>
<tr>
<th>MUTATION</th>
<th>PHENOTYPE</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>G973S</td>
<td>severe</td>
<td>Gomez-Lira et al. (1993) Fifth Int. Conf. on OI abst. 120</td>
</tr>
<tr>
<td>G973V</td>
<td>lethal</td>
<td>Lamande et al. (1989) J. Biol. Chem. 264: 15809–15812</td>
</tr>
<tr>
<td>G976R</td>
<td></td>
<td>Lamande et al. (1989) J. Biol. Chem. 264: 15809–15812</td>
</tr>
<tr>
<td>G1009S</td>
<td>severe</td>
<td>Cohn et al. (1990) Matrix 10: 236 (abst.)</td>
</tr>
<tr>
<td>G1009V</td>
<td>lethal</td>
<td>Cohn et al. (1990) Matrix 10: 236 (abst.)</td>
</tr>
</tbody>
</table>
Table 1.4 Glycine substitutions in the COL1A2 gene

Glycine substitutions characterised in the COL1A2 gene are listed with the resulting phenotype. Unless the mutations are discussed further in the text, references are given in shortened form and are not included in the REFERENCES section. Mutations are indicated by the glycine residue number followed by the substituting residue. Amino acids are indicated by the standard single letter code. Where no phenotype is given it is the same as the preceding one.

<table>
<thead>
<tr>
<th>MUTATION</th>
<th>PHENOTYPE</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>G247C</td>
<td>severe</td>
<td>Marini et al. (1993) Fifth Int. Conf. on OI abst. 126</td>
</tr>
<tr>
<td>G496R</td>
<td>lethal</td>
<td>Bateman et al. (1990) Fourth Int. Conf. on OI abst. 2</td>
</tr>
<tr>
<td>G661S</td>
<td>lethal</td>
<td>Gomez-Lira et al. (1993) Fifth Int. Conf. on OI abst. 120</td>
</tr>
<tr>
<td>G696R</td>
<td>lethal</td>
<td>Gomez-Lira et al. (1993) Fifth Int. Conf. on OI abst. 120</td>
</tr>
<tr>
<td>G706S</td>
<td>lethal</td>
<td>Grange et al. (1990) Nucl. Acids Res. 18: 4227–4236</td>
</tr>
<tr>
<td>G865S</td>
<td>lethal</td>
<td>Lamande et al. (1985) J. Biol. Chem. 264: 15809–15812</td>
</tr>
<tr>
<td>G1012R</td>
<td>moderate</td>
<td>Wenstrup et al. (1991) J. Biol. Chem. 266: 2590–2594</td>
</tr>
<tr>
<td>G976D</td>
<td>lethal</td>
<td>Byers (1990) Trends Genet. 6: 293–300</td>
</tr>
<tr>
<td>G1012R</td>
<td>moderate</td>
<td>Wenstrup et al. (1991) J. Biol. Chem. 266: 2590–2594</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>-----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>CYS</td>
<td>O</td>
<td>Mi</td>
</tr>
<tr>
<td>ASP</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>SER</td>
<td></td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>Mi</td>
<td>L</td>
</tr>
<tr>
<td>VAL</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>ARG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.1 Glycine substitutions in the triple helical domain of the α1(I) collagen chain.

The positions of various amino acids substituted for glycine along the length of the triple helical region of the α1(I) chain are illustrated. The resulting OI phenotype is indicated below the position. L, lethal; S, severe; Mo, moderate; Mi, mild; Os, osteoporosis. References for the above mutations can be found in table 1.3.
<table>
<thead>
<tr>
<th></th>
<th>CYS</th>
<th>ASP</th>
<th>SER</th>
<th>VAL</th>
<th>ARG</th>
<th>ALA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>247 259</td>
<td>L</td>
<td>567 580</td>
<td>L 708</td>
<td>805</td>
<td>907</td>
<td>976</td>
</tr>
<tr>
<td>640 666</td>
<td>L MaS</td>
<td>661</td>
<td>906</td>
<td>865</td>
<td>922</td>
<td></td>
</tr>
<tr>
<td>787</td>
<td>L</td>
<td>O L MaS</td>
<td>L Va</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>457 496</td>
<td></td>
<td>544 586</td>
<td>676</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.2** Glycine substitutions in the triple helical domain of the α2(I) collagen chain

The positions of various amino acids substituted for glycine along the length of the triple helical region of the α2(I) chain are illustrated. The resulting phenotype is indicated below each position. L, lethal; S, severe; Mo, moderate; O, osteoporosis; Va, variable. References for the above mutations can be found in table 1.4.
GGG codon is the rarest occurring of the glycine codons. It represents just 3.3% of all glycine residues in the first position of the Gly-Xaa-Yaa repeat motif in the α1 mRNA and just 4.8% in the α2 mRNA of the triple helical domain of type I collagen. Substitutions by alanine and valine (other neutral, hydrophobic amino acids that can be produced following a point mutation in a GGN codon) result in the lethal phenotype when they occur in the α1 chain and the two glycine to valine substitutions in the α2 chain result in non-lethal phenotypes; glycine by alanine mutations in the α2 chain have not been reported. The more severe phenotype arising from mutations in the proα1 chain as opposed to those in the proα2 chain may be due to the greater proportion of mature collagen molecules being defective or an α2 mRNA domain effect. From this information it could be predicted that glycine by tryptophan substitutions might be lethal if they occur in the proα1 chain. However, since there are no known glycine to tryptophan mutations, it is not possible to predict with any certainty the phenotypic consequences of such mutations. Similar arguments can be invoked to explain why only one glycine by glutamic acid substitution has been reported (Rose et al., 1993); this mutation is discussed in detail in Chapter Seven.

There are four possible codons for alanine (GAN) which would all result from a G to A transition in the second position of the GGN codon. There are also four encoding valine (GTN) which would result from a G to T transversion in the second position of a glycine codon. Thus it would seem that both amino acids have an equally likely chance of being substituted. However, seven glycine to valine substitutions in type I collagen have been reported as opposed to the two alanine substitutions reported. The G to A transition is not that uncommon in the second position as indicated by the number of aspartic acid substitutions reported — 11 in the two type I α-chains. Therefore there must be another reason for the paucity of alanine substitutions. It is possible that the lethal phenotype resulting from the α1 substitutions is an exception; they may just happen to occur in a domain which promotes such a phenotype. It may be that other glycine by alanine substitutions exist but result in a mild phenotype which may go undiagnosed if the affected individual encounters relatively few problems, such as minor or infrequent fractures.

The most readily characterised of the substitutions is that by cysteine. Since it is an amino acid normally excluded from the triple helical domain of type I collagen chains, it is easily detected at the protein level due to the formation of disulphide-bonded α-chain dimers. The substitutions by both cysteine and serine illustrate the effect that domains within the triple helical region of the collagen molecule might have on the resulting phenotype since the phenotypic outcome of these substitutions does not adhere to a perfect gradient effect of substitutions from the carboxyl- to the amino end of the molecule (in figures 1.1 and 1.2). The mutation at residue α1(1)1009 which is the most carboxyl terminal of all the reported serine substitutions results in a severe but non-lethal phenotype. This may reflect a domain effect or a more complex mutational effect of this region. Substitutions by arginine appear to conform to a gradient effect of lethal and severe OI phenotypes and glycine by aspartic acid or valine substitutions all result in a lethal phenotype.

Fewer mutations in the α2(1) mRNA leading to glycine substitutions in the mature α-chain have been identified. Thus the existence of phenotypic gradients and possible domain arrangements is more difficult to predict. However, with the identification of more mutations, this may be facilitated. Currently, most of the mutations found result in a phenotypic gradient comparable to that seen in the α1(1) chain, the exception being the...
glycine by serine substitutions. However, as can be observed from figure 1.2, there is a dearth of reported mutations at the N-terminal end of the α2(I) chain. It is highly unlikely that mutations in this region do not exist, rather that the phenotype may be sufficiently sub-clinical to have gone undetected.

There is one possible mutation of a single glycine codon that could result in a stop codon being substituted, namely GGA to TGA. Once again, the relative infrequency of the codon for the substituted residue may result in this being a rare event. The effect of stop codons in the coding sequence is discussed in section 1.5.1.4.

In general, the mutations in either type I collagen chain giving rise to an OI phenotype are 'private', i.e. restricted to an individual or a family. However, it is becoming apparent that unrelated individuals having a common mutation do exist. In some of these cases, the phenotypes of the individuals identified differ in severity. The possibility of an additional defect in another gene causing the more severe phenotype may be likely. Examples of such multiple occurrences of mutations are detailed in Chapters Four, Five and Nine.

(b) Exon-skipping mutations

Second to point mutations, exon-skipping mutations in type I collagen are the most common cause of the OI phenotype. In general, exon-skipping arises due to point mutations in the vital consensus splice donor or acceptor sites or small genomic deletions. To date, mutations causing the skipping of non-triple-helical exons have not been documented although several have been identified within the helical region. Usually the phenotype is very severe or lethal (Bonadio et al., 1990; Wallis et al., 1990a) presumably due to the enormous disruptive effect on the forming helix, and is essentially comparable to a large deletion. However, mild phenotypic effects have been observed resulting from the skipping of α1(I) exon 8 (Bateman et al., 1990) and α2(I) exon 16 (Filie et al., 1993) in which the majority of the helix has been formed already (α1(I)) such that chain abnormality is minimised, or occurs in the minor species α-chain (α2(I)) such that the effect on the abundance of the mature molecule is minimised. It should be noted that the skipping of α1(I) exon 14 produces a lethal phenotype — this individual was homozygous for the mutation (Bonadio et al., 1990). Mild phenotypes have also been found to be the result of a mutation which only affects a proportion of mature collagen molecules, as has been observed from the skipping of α1(I) exon 17 (Pruchno et al., 1989). Within the COL1A2 gene, a 5' – 3' gradient of increasing severity of phenotype exists. A number of skipping mutations have been collated by Byers (1993), namely that of α2(I) exon 12 resulting in type I OI, α2(I) exons 11, 13, 21 and 26 resulting in type IV OI, α1(I) exons 14, 44, 47, α2(I) exons 28 and 33 resulting in OI type II.

In both COL1A1 and COL1A2, the skipping of exon 6 results in an Ehlers-Danlos syndrome type VII phenotype (section 1.5.3) due to the exclusion of the amino-terminal procollagen protease site from the procollagen molecules which are, as a consequence, incorrectly processed.

(c) Multixon deletions

Such rearrangements are rare despite the repetitive structure of type I collagen genes; possible explanations for this have been discussed in section 1.4.4. The phenotypes resulting from such mutations have all
been lethal presumably due to the very disruptive effect they have on helix formation and the high degree of protein degradation. A deletion of exons 23–25 of the COL1A1 triple helical domain (Chu et al., 1985) has been identified and a seven exon deletion (Δ34–40) within the triple helical domain of COL1A2 (Willing et al., 1988). The deletion of α1(I) exons 27–29 (Barsh et al., 1985) has also been characterised.

(d) Short deletions

Several short deletions have been reported and are often the result of splicing defects. An 11 base pair deletion from an intron in the COL1A2 gene caused the skipping of exon 9 and mild atypical OI in one patient (Nicholls et al., 1992). The phenotype was characterised by excessive joint hypermobility and premature osteoporosis leading to late-onset fractures. Kuivaniemi et al. (1988a) reported a 19 base pair deletion which resulted in the skipping of exon 11 in the proband whose family displayed atypical OI possibly due to the differential exon skipping efficiency in different family members. Wallis et al. (1992) and Hawkins et al. (1991) reported two independent Gly-Ala-Pro triplet deletions of the α1(I) chain both of which resulted in lethal OI. The first was a deletion of residues 874–876 and the second deleted one of three consecutive GAP triplets at positions 868–876. Another triplet deletion (residues 732–734) was identified in the α1 chain (Byers, 1993). The mechanism by which such inframe deletions cause a lethal phenotype is unclear but was suggested that it may be due to abnormal chain registration (Hawkins et al., 1991). This suggestion was based on the idea that immediately prior to the α-chains adopting a helical conformation, the three chains align in a ‘loose conformation’, i.e. chain registration. It was proposed that this tripeptide deletion might interfere with the chain association, potentially due to the different Xaa and Yaa position amino acids now present at each position N-terminal to the deletion. This presumably delays helix formation and results in over-modification even though the mutation does not disrupt the perfectly repeating Gly-Xaa-Yaa motifs. The deletion of the triplet from the repeated region suggests that repetitive regions may be involved in DNA rearrangements particularly if they have a high GC content (Chandley, 1989). A Gly-Pro-Pro deletion from residues 1009 to 1011 in the α2(I) chain (Byers et al., 1991) resulted in a mild phenotype. This could be a result of the mutation not occurring in a critical region of the C-terminus, of it being in the α2 chain and thus a loss of only 50% of all collagen molecules may occur (section 1.5.1.4) or the mutation affecting the rate of collagen biosynthesis resulting in a decreased amount of protein in the ECM.

Smaller deletions have been detected; a single triplet encoding valine at position 255 in the α2(I) chain was found to be deleted in a proband with type III OI (Molyneux et al., 1993). The phenotype was probably caused by the degradation of secreted abnormal collagen molecules. A five base pair deletion towards the 3’ end of COL1A1 produces a shift in the reading frame which was predicted to result in a molecule 84 residues longer than normal (Willing et al., 1990). It was suggested that the mutant allele was processed and available for translation but might have been rapidly degraded and therefore not incorporated into a collagen molecule. The effect of this ‘null’ allele was a type I phenotype (section 1.5.1.4).
Compared to the types of mutation discussed above, there are comparatively few insertional mutations documented. An insertion of 600bp into the α1(I) mRNA resulting from a recombination event between exons 14 and 17 caused lethal OI (Byers et al., 1988a) as did a tandem duplication of exons 15 and 16 in the COL1A1 gene (Cohn et al., 1993) and an insertion of 75bp from intron 35 into the α1(I) mRNA (Genovese et al., 1989). Although the latter would preserve the open reading frame, it is possible that due to the size of the insert, the mature molecule may possess a loop structure which interferes with secretion of the molecule or involvement in fibril formation. It could be inferred from such an argument that smaller insertions that maintain the open reading frame would cause a less severe phenotype because of the smaller loop formed. A point mutation in COL1A2 which alters a consensus splice site from GT to TT in a patient with moderate OI may substantiate this argument. This base alteration resulted in the use of a second GT dinucleotide and caused an insertion of only six residues in the transcribed sequence (Byers, 1993). Another six residue insertion was the result of a single base substitution α2(I) in intron 33 resulting in the use of an alternative splice site. The individual suffered from type IV OI (Wenstrup et al., 1993).

An example of a one base insertion is that of a single thymine after nucleotide 4088 of exon 49 in the preproα1(I) mRNA (non-triple-helical region) in an individual which resulted in a truncated carboxy-terminal end of the proα1(I) chain. There appeared to be complete degradation of the mutant molecules and only normal type I collagen was detected in the ECM at approximately 20% of the normal amount as compared to control tissues. The infant died from OI type IIIB (Cole et al., 1990).

1.5.1.2 Overmodification and locality of mutation

If a glycine residue is replaced by a bulkier amino acid, helix winding and modification proceed as normal up to the point of substitution. Since the new amino acid is bulkier than a glycine, it takes longer to be incorporated into the middle of the growing helix, thus helix winding is delayed momentarily. This results in the unwound α-chains acting as substrates for the modifying enzymes for a greater length of time and thus they become overmodified. Once the substituted residue has been incorporated, helix propagation continues as normal. The proα-chains are over-modified from the position of mutation to the N-terminus of the polypeptide. This over-modification is present in each of the three component α-chains regardless of the chain in which the substitution resides. Over-modification has been shown to occur as a result of single amino acid deletions (Molyneux et al., 1993) as expected, but also as a result of a 9bp (Gly-Ala-Pro) deletion even though the chains remained in correct register (Hawkins et al., 1991; Wallis et al., 1992).

The extent of over-modification is not necessarily the primary determinant of the OI phenotype. It may be considered a secondary effect of the mutation on the collagen molecule and as such has been useful in determining the approximate location of the mutation within the collagen molecule. The collagen polypeptides can be fragmented by digestion with cyanogen bromide which cleaves C-terminal to methionine residues to yield several smaller peptides of known size (Barsh and Byers, 1981; Bonadio et al., 1985). On an SDS-polyacrylamide gel, the mobility of an overmodified peptide fragment is altered with respect to a normal one. From this analysis, the peptide fragment which harbours the mutation can be predicted. The fragments generated from the digestion of
the α(I) chains are of varying lengths. In comparison, those generated from a similar digestion of the α2(I) chain are less heterogeneous in length and are more difficult to separate efficiently by gel electrophoresis. Thus results from such an analysis are usually reported with respect to an α1(I) cyanogen bromide (CB) peptide fragment even though the mutation may occur in the α2(I) chain.

1.5.1.3 Mosaicism

Mosaicism can be either gonadal or somatic. If an individual possesses two different types of an allele within the germ cells, they are referred to as a gonadal mosaic. When an individual is a true gonadal mosaic, they show no degree of phenotype associated with the defective allele but since the allele is present in their germ cells, their children may display the phenotype. There is no means of telling if an individual is mosaic prior to the birth of the first affected child and the mutational event may occur at any time, e.g. in the germ cells once they have differentiated or later in life due to environmental factors. A somatic mosaic individual arises when the mutation occurs during embryogenesis; such an individual can also be a gonadal mosaic. The mosaic person may or may not exhibit symptoms associated with the mutation depending on the degree and nature of the affected cells; offspring from a gonadal/somatic individual are likely to be more severely affected than the parent.

Mosaicism for mutations has been recognised in OI as being the cause of recurrence within a pedigree. In families with multiple OI-affected siblings, de novo mutations are statistically unlikely to be the cause of recurrence. The recurrence risk for OI type II is estimated to be 6−7% (Byers et al., 1988b) which is a figure below that expected for autosomal recessively inherited mutations. Parental germline mosaicism for the gene defect was proposed as the cause (Cohn et al., 1990a, c; Edwards et al., 1990). Germine mosaicism means that a parent harbours both normal and abnormal alleles in the germ cells and these can be inherited by more than one offspring. In general, the parents are asymptomatic although Constantinou et al. (1990a) reported a case in which the parent displays mild OI symptoms. This was explained by the fact that the mutation in the mother must have occurred de novo prior to cell differentiation in embryogenesis such that both her somatic and germline cells harboured the mutant allele. Constantinou-Deltas et al. (1993) reported a similar phenomenon. The degree of mosaicism in the parent may also cause OI symptoms (Edwards et al., 1990; Wallis et al., 1990b).

Mosaicism has also been implicated in type III OI (Abuelo and Byers, 1991). A possible explanation for the recurrence in such cases could be the differential expression of the mutant gene.

1.5.1.4 Biochemical consequences of OI mutations

The mutations discussed above can either cause the defective α-chain to be excluded from or included in the procollagen molecule. From the data available, generalisations can be made on the severity of phenotype expected from certain types of mutations.

An α-chain can be excluded from a molecule by one of two mechanisms, (i) the failure of an allele to yield a product, or (ii) the failure of a molecule to incorporate a defective polypeptide. In both cases, the resulting phenotypes tend to be milder than those resulting from the presence of a defective molecule in the matrix tending to result in OI type I. Mutations which destroy sequences vital for transcription of the gene (e.g. promoter or enhancer
regions) would yield no transcript and therefore there would be a lower number of collagen α-chains for processing into mature collagen resulting in a decrease in the amount of collagen in the ECM. The lower abundance or absence of an α-chain could also be explained by the deletion of a complete allele, introduction of a termination codon and splicing mutations (both potentially resulting in a shortened transcript which may or may not enter into the cytoplasm). This so called 'null allele' hypothesis can explain the occurrence of mild OI (reviewed by Byers, 1993; Willing et al., 1992). Mutations that destroy sequences vital for chain assembly would cause the defective chain to be excluded from the helix and result in the same phenotype. The region of importance for initiation of helix formation is the C-terminal propeptide. A glycine by cysteine substitution at position 1017 in the α2(I) chain (Cohn et al., 1988) resulted in a mild phenotype. However, a four base pair deletion in the C-terminus which caused a frame shift in the α2(I) chain (Pihlajaniemi et al. 1984) caused moderate OI; no α2(I) chains were synthesised and an α1(I) homotrimer was the only type I collagen species present. The existence of a stop codon introduced into the C-terminal propeptide resulting from a single base mutation has not been reported. Such a mutation would also be expected to result in exclusion of the α-chain from the helix. Stover et al. (1993) reported an aberrant splice site in intron 26 of one COL1A1 allele. The retained intron introduced an inframe stop codon and an out of frame insertion within the mRNA. This resulted in additional stop codons further downstream and the failure of the mRNA to be processed. The individual displayed type 1 OI.

Other mutations allow the incorporation of the defective α-chain into the helix. These tend to have more deleterious effects. For these more deforming types of OI, the phenotype appears to reflect the chain in which the mutation occurs, the location of the mutation in the chain and the nature of the mutation itself (section 1.5.1.1a).

In general, for point mutations, a milder phenotype occurs if the mutation is further N-terminal in a chain. This is consistent with the proportion of the chain which may be defective and the degree of overmodification compared to that resulting from mutations further C-terminal. Depending on the type of mutation, the register of the chains N-terminal to the mutation may be disrupted which might prevent interactions between molecules that would normally stabilise the helix. It may permit excessive overmodification N-terminal to the mutation thus generating an asymmetric molecule, or might delay the secretion of one or more species of abnormal molecules. Such abnormal molecules could be poorly processed extracellularly and thus may interfere with normal fibril assembly.

Some mechanisms can amplify the effect of a mutation. When a bulkier amino acid occurs in place of a glycine within the triple helix or a small deletion occurs, the overmodification that takes place often has an effect on the thermal stability of the collagen. This alteration to thermal stability may be assayed in vitro by the altered susceptibility to degradation by trypsin. The effects have been reviewed by Bichinger et al. (1993). Mutations causing OI decrease the denaturation temperature often by 2 to 4°C but occasionally a more dramatic decrease is seen (e.g. the 21°C decrease observed by Westerhausen et al., 1990). A severe lowering of the denaturation temperature tends to prevent the folding of the polypeptides at bodily temperatures and thus a procollagen is not secreted into the extracellular space. Procollagens with markedly reduced helical formation accumulate in the fibroblasts and are degraded by non-specific proteinases as they slowly pass out of the cells. One single mutated α-chain can prevent the folding of a molecule that contains both a normal and a mutant chain, thus the effects of the
mutation are amplified. If the mutation occurs in the α1(I) chain two possible molecules containing one defective α-chain and one with two defective chains exist. Thus if all molecules containing at least one defective chain are degraded, only 25% of total collagen expected will be present in the cells e.g. the frameshift mutation reported by Bateman et al. (1989). This can be compared to 50% of total collagen if the mutation occurs in the α2(I) chain. The term ‘procollagen suicide’ has been coined and serves to explain such observations (Prockop, 1984). The second amplification mechanism results from the production of a "kinked" collagen molecule or fibril. A molecule with a flexible kink was identified by Vogel et al. (1987, 1988) and resulted from a glycine to cysteine change at position 748 in the α1(I) chain. It was predicted by model building and confirmed by rotary shadowing EM that the cysteine was accommodated in a loop and that, amino-terminal to the point of substitution, the α-chains were out of phase. This altered phase of the chains altered the conformation of the N-proteinase cleavage site. The proband’s procollagen was more resistant to cleavage by N-proteinase compared to the normal procollagen molecules. It was subsequently shown that a mixture of normal and kinked molecules generated highly branched and dendritic fibrils not seen in a normal individual. Thus the morphology of the fibrils was altered and resulted in a lethal phenotype. Recent work has shown that not all substitutions C-terminal to the N-proteinase cleavage site that result in the reduction of cleavage by the enzyme introduce a kink in the molecule (Lightfoot et al., 1992). In fact, only the substitution by cysteine appeared to cause such a kink when compared to substitutions by aspartate and arginine which suggests that a kink is not necessary for the altered cleavage rate of the procollagen by N-proteinase.

The process of overmodification can result in an asymmetric molecule which may have an effect on efficient fibril formation.

1.5.1.5 Social issues and OI

The pre-natal diagnosis of OI is hampered by the observation that the majority of mutations are ‘private’, i.e. are restricted to an individual or a family and that the disorder is heterogeneous in nature. Type III OI is the progressively deforming type. However, the severity varies from individual to individual such that some older individuals may exhibit a phenotype no more severe than that of other unrelated, younger individuals with the same type of OI. Only a few cases have been reported in which the same mutation was found in unrelated individuals. Pre-natal diagnosis of OI often relies on a non-invasive approach such as radiography and ultrasonography which are usually carried out as a procedure in a pregnancy. The use of such methods has been reviewed (Thompson, 1993). Such methods are useful mainly for the diagnosis of severe forms of OI, in particular type III, since the milder forms may go undetected due to the lesser extent of deformity. Such diagnoses can prevent unnecessary Caesarean section operations if the infant is expected to have a limited life expectancy. Invasive testing is usually carried out to identify affected individuals in families with a known history of OI in which the mutation has been identified.

In post-natal diagnoses, it is a recognised problem that unexplained fractures which are characteristic of OI can be mistaken for non-accidental injury (NAI) in some instances. The sporadic nature of many OI mutations and the absence of other disorder characteristics such as blue sclerae, decreased bone density and bone malformation, can lead to a diagnosis of child abuse and possible removal of the child into care until a correct
diagnosis is made (Paterson and McAllion, 1989; Paterson et al., 1993). The recognition of a possible variant form of OI in which a temporary period of fractures occur in the first year of life (Paterson et al., 1993) no doubt serves to confuse diagnosis further.

Currently there is no cure for OI and no effective medical treatment. Numerous trials have been carried out, e.g. the administration of somatotrophin by Morley et al. (1993) but benefits are as yet unknown and may be limited. The management of OI is directed at the treatment of fractures promptly to maintain mobility and to prevent osseous deformity, the use of intramedullary rodding of the long bones, the reduction or prevention of scoliosis, and surgery to correct hearing loss. For severe cases, external support for bones can be applied and can aid standing and walking. For many individuals with severe survivable OI, wheelchairs may be the only appropriate mode of mobility. Care for the few short-term survival cases of type II OI has involved intensive care in handling, increased prevention of contact with pathogens and possibly respiratory assistance.

1.5.2 Osteoporosis

Osteoporosis is defined as the degree of skeletal loss at which bone is unable to withstand ordinary stress without fracturing. Pathologically, osteoporosis is characterized by an absolute decrease in the amount of bone. Thus it can be seen that from a clinical angle, there appears to be a phenotypic overlap with mild (type I) OI. The disorder is heterogeneous in nature and although it is thought to have many causes, it often appears to be familial (reviewed by Riggs and Melton, 1986).

In the main, two populations of individuals are affected with osteoporosis. Post-menopausal women form the first group and often suffer from the so-called type I osteoporosis which is likely caused by a decrease in oestrogen levels which can sometimes be corrected by the administration of this hormone (Felson et al., 1993). The second group comprises both men and women over 70 years of age; type II osteoporosis. Mutations similar to those observed in OI-affected individuals, namely glycine substitutions (α1(I)G19C and α1(I)G43C (Shapiro et al., 1992); α2(I)G661S (Spotila et al., 1991)) and exon skipping events (skipping of α1(I) exon 9 (Nicholls et al., 1992)) in type I collagen genes, have been reported in a few osteoporotic individuals. Thus it can be seen that mutations in type I collagen genes can result in OI-like, but separately classified, disorders.

1.5.3 Other disorders

Mutations in type I collagen genes have also been found to be a cause of Ehlers-Danlos syndrome (EDS) type VII. The EDS syndromes form a heterogeneous group of disorders with approximately eleven clinically different types identified to date as discussed below (Beighton et al., 1988). Aspects of all types of EDS have been reviewed by Steinmann et al. (1993). The general features of EDS are extreme joint hypermobility and hyperviscosity of skin, tissue fragility, cigarette-paper like scarring of the skin and easy bruising. Type VII is additionally characterized by congenital bilateral hip displacement and muscular hypotonia. In all patients for whom a molecular defect has been identified, exon 6 skipping of either type I collagen α-chain appears to be the underlying cause. The skipping of the 24 (α1) or 18 (α2) amino acids destroys the N-proteinase cleavage site, a lysine residue involved in cross-linking and other peptide cleavage sites (Steinmann et al., 1993). The retention
of the N-propeptide causes the molecule to be more soluble than normal which may be significant in the pathology of the disorder. Normal fibril formation is probably inhibited and any fibrils formed will be under cross-linked. There have been several mutations characterised and type VII has been sub-grouped according to the type of mutation involved (Beighton et al., 1988). The incomplete processing of the N-propeptide is another mechanism whereby the effect of a mutation can be amplified. Two types of mutation can affect the processing by N-proteinase, namely (a) in-frame deletion of an amino acid sequence which is important in the cleavage site of the enzyme, and (b) in-frame deletion of an amino acid or an amino acid change far from the cleavage site that either reduces or prevents cleavage of the procollagen because of the long range effect of the mutation on the cleavage site. Some of these mutations have been reviewed by Prockop et al. (1990) and recently Watson et al. (1992) reported another instance of incomplete cleavage by N-proteinase resulting from a G to A transition in intron 6 at the 5’ donor splice site such that half the proc2(I) mRNA lacked exon 6. At the cleavage site, each of the three α-chains forms a hairpin loop structure. Mutations which occur further from the cleavage site may have an effect on this domain through incorrect chain assembly at the critical point, again resulting in the failure, or reduced rate of cleavage by the enzyme (e.g. Vogel et al., 1988; Lightfoot et al., 1992; Watson et al., 1992). Such mutations can often cause all three chains to be resistant to cleavage or, if the molecules are partially cleaved, the presence of partially processed procollagens in a mixture of normal collagens can cause the formation of thin and irregular fibrils (reviewed by Prockop et al., 1990).

Mutations in type II collagen have been implicated as the underlying defect in some forms of chondrodysplasia, for example, achondrogenesis and the Stickeker syndrome, and osteoarthritis. The chondrodysplasias are a heterogeneous group of heritable cartilage disorders characterised by dwarfism, joint deformities and other skeletal defects. Mutations comparable to those found in OI have been found (reviewed by Kuivaniemi et al., 1991). Linkage to type II collagen was demonstrated in families with osteoarthritis, a disorder characterised by progressive degeneration of the joint cartilage (Palotie et al., 1989; Knowlton et al., 1990). Other mutations characterised in the COL2A1 gene have included a base change causing a glycine to serine substitution in the α-chain (Vissing et al., 1989) and an A579C substitution (Ala-Kokko et al., 1990). A 45bp insertion into exon 48 resulting from an internal tandem duplication causes spondyloepiphyseal dysplasia in the patient (Tiller et al., 1990) and a deletion in the same exon (Lee et al., 1989) results in achondrogenesis, characterised by dwarfism. The Stickeker syndrome affects the eyes causing severe myopia, degeneration of the retina and vitreous humour. Since type II collagen is not detected in the retina itself (Marshall et al., 1993) it can be postulated that the effect of the mutation on the composition of the vitreous humour causes the accompanying retinal detachment. The disorder is often associated with moderately severe osteoarthritis or a chondrodysplasia. A mutation that substituted a stop codon for an arginine codon has been identified in an affected family (Ahmad et al., 1990).

Type III collagen is often found in association with type I but is absent in bone and tendon and occurs in abundance in vessel walls, skin and gut. Mutations in type III collagen include a number that have been shown to be responsible for Ehlers-Danlos syndrome type IV (see above and Kuivaniemi et al., 1991) which is the most severe form; in addition to the main features of EDS, type IV is characterised by the sudden rupture of large arteries or intestines, in particular the bowel. The mutations identified in patients include single base changes
converting glycine codons to codons for bulkier amino acids (as in osteogenesis imperfecta), RNA splicing mutations, and gross deletions (reviewed by Kuivaniemi et al., 1991; Steinmann et al., 1993). It appears that the generalisations made with respect to 01 mutations and phenotypes can also be extended to those found in patients with EDS IV. Individuals with arterial aneurysms have been reported in which a type III collagen mutation has been detected (Kontusaari et al., 1990a, b). Type III collagen abnormalities have also been detected in patients with cerebral aneurysms (Majasni et al., 1992). Abdominal aortic aneurysms appear to be familial and thus may be caused by genetic defects, however, elastin and fibrillin are major proteins also present in arteries and thus the possibility of defects in the corresponding genes being a cause of the disorder cannot be eliminated on evidence reported to date. A phenotypic overlap of mild type IV EDS and AAA can be seen (Powell et al., 1991).

Mutations in type IV collagen have been identified in individuals with Alport’s syndrome, an X-linked nephritis associated with hearing loss. This disorder has been reviewed by Tryggvason et al. (1993) and is characterised by the splitting of the glomerular basement membrane in the kidneys leading to renal failure. Several different mutations in the α5(IV) collagen gene have been reported. These include a partial (15kb) gene deletion at the 3’ end (Barker et al., 1990) which results in the loss of a functional α5 chain (Hostikka et al., 1990; Zhou et al., 1991a, b), a gene rearrangement (Boye et al., 1991) and an α5(IV) G325A substitution (Knebelmann et al., 1992). It has been suggested that mutations in the COL6A3 gene may cause oesophageal leiomyomatosis (Zhou et al., 1993).

Type VII collagen is the major component of the anchoring fibrils which act as attachment structures keeping the papillary dermis and basement membrane together. Recently a mutation in the type VII collagen gene has been shown to be responsible for the recessively inherited form of dystrophic epidermolysis bullosa (Christiano et al., 1993). This heterogeneous disorder is characterised by blistering of the skin as a result of minor trauma; different forms of the disorder are classed according to the degree of tissue separation. In the dystrophic type the blister forms within the papillary dermis and below the basement membrane. Genetic linkage between the COL7A1 locus and both the recessive and dominant inherited forms of the disorder have been shown (Ryynanen et al., 1991; Christiano et al., 1993).

A deletion of 13 bases in one type X collagen allele may be the cause of autosomal dominant Schmid metaphyseal chondrodysplasia (Warnan et al., 1993). The mutation introduces a frame-shift in the C-terminal domain known to be important in chain association. It is speculated that the mutation results in a reduced amount of type X collagen in the hypertrophic chondrocytes. The disorder is characterised by short, but non-dwarf, stature, rib defects and metaphyseal involvement.

It has also been suggested that the nail-patella syndrome (NPSI) could be due to a defect in the COL5A1 locus since this gene has been chromosomally located to a region near the locus for NPSI (Greenspan et al., 1992). The disorder is characterised by dysplastic nails and absent patellae. A proband with a deletion in the COL5A1 gene has recently been identified as having NPSI-like pathology (Ghiggeri et al., 1993).
1.5.4 The use of mutant animals in the study of osteogenesis imperfecta and other disorders linked to collagen gene defects

Syndromes that have been noted to phenotypically simulate OI have been reported in cattle, cats, tigers and mice (Chipman et al., 1993). Transgenic mice have been used in various instances in order to study the mutations in collagen genes underlying the phenotype and for discovering additional diseases that may be caused by collagen defects.

The fortuitous insertion of Moloney murine leukaemia virus (M-MuLV) into the Mov locus of mice resulted in a lethal recessive mutation causing homozygotes to die in utero (Jaenisch et al., 1983; Harbers et al., 1984). It was subsequently shown that the locus corresponded to the murine α1(I) collagen gene and the insertion occurred in the first intron of the gene. The time of death (11–12 day gestation) corresponds to the time at which elevated transcription of the COL1A1 gene occurs in the normal embryo. It was recognised that the mutation destroys a site with a known involvement in transcriptional regulation (section 1.4.5.1).

Transgenic mice were produced which harboured either a G859C or a G859A substitution in the α1(I) collagen chain (Stacey et al., 1988). Both mutations exerted a dominant lethal effect and the mice presented with a pathology and biochemistry comparable with that seen in human type I I OI. There are no directly comparable mutations in humans although a G859 substitution has been found (Chapter Four). However it is on the α2(I) chain which might explain the less severe phenotype (type III) than that seen in the mice, and it is substituted by serine and therefore comparisons are perhaps not entirely valid. A mini α1(I) gene which lacked 41 central exons (Olsen et al., 1990) was shown to simulate procollagen suicide in a lethal form of OI in transgenic mice (Khillan et al., 1991).

The freifro mouse (fragilitas ossium; Guenet et al., 1981) was identified as having a phenotype similar to OI. This mouse has been demonstrated to have clinical and pathological similarities with recessively inherited OI types II and III (Sillence et al., 1993). All of the mice have the same mutation, however, 90% of cases resulted in perinatal death whereas the remainder were survivors. It is possible that the mutation in these mice may indicate the border between the two OI phenotypes. Clearly something other than the mutation plays a role in the overall phenotype and it was suggested that lethality may not be the best feature on which to base the characterisation (Sillence et al., 1993).

Chipman et al. (1993) reported a naturally occurring mouse (oim) which displayed a phenotype similar to moderate to severe OI. They noted that the mice had a progressively deforming type of the disorder but with no notable premature mortality. On examination, the gene defect was found to be the deletion of a G at nucleotide position 3983 (exon 52) in the murine cola-2 gene (analogous to the human COL1A2 gene). This mutation resulted in the alteration of the last 48 amino acids in the murine α2(I) collagen chain which were subsequently prevented from becoming incorporated into a triple helix with α1 chains. As a result of this absence of α2 chains, an α1(I) homotrimeric collagen was seen to accumulate in the tissues studied, namely skin, bone and dermal fibroblasts. There is no degradation of the collagen molecules since the defective chain is not incorporated into the procollagen molecule. The mode of inheritance of the mutation was determined as recessive. The thermal stability of the oim collagen is similar to that reported for a human α1(D)3 molecule and the presence of such a molecule is implicated.
in the destruction of multiple components of the ECM such as reduction of proteoglycan, and the lateral packing of
the molecules (McBride et al., 1993). The transgenic mouse of Chipman et al. (1993) cannot be considered to be a
model for the de novo incidences of human type III OI which is the predominant mode of occurrence, however, it
may provide insight into the clinical and pathological features of this sub-group of OI.

Other examples of mouse models for OI include inbred lines of mice expressing a mutant COL1A1 or
COL1A2 gene with resultant mild phenotypes mirroring the phenotype of human osteoporosis (Pereira et al., 1993)
or osteoarthrosis (Helminen et al., 1993). Various workers (Garcia et al., 1991; Vandenberg et al., 1991;
Metsiranta et al., 1992) have reported mice expressing a heritable mutant COL2A1 gene which results in a
phenotype similar to human chondrodysplasias.

Transgenic mice can also be used for the detection of other collagen disorders for which no defect has
been discovered in humans. An example of such a case is a mouse possessing a partial deletion of the murine
equivalent of the COL9A1 gene. The animal has a phenotype analogous to osteoarthrosis with a mild
chondrodysplasia (Kivirikko, 1993) suggesting that mutations in the human type IX collagen genes may be
responsible for the corresponding disorders in man as well as those found in type II collagen genes. Yet another
example is a mouse synthesising a truncated form of the murine type X collagen which developed
spondyloepiphysial dysplasia; human type X collagen gene defects may result in similar human phenotypes as
observed by Warman et al. (1993).

1.6 THE DETECTION OF MUTATIONS

If a protein product is unknown the mutant locus must be identified (a so-called reverse genetics
approach). This process starts by the identification of a disease phenotype and then relies on genetic mapping
techniques to identify the chromosomal location of the mutant locus. The isolation of clones containing
overlapping regions of sequence can identify putative gene sequence and the gene of interest can be located.
Sequencing then identifies the nature of the gene and a protein product can be identified. Genetic mapping of a
gene locus proceeds by linkage analysis; pairs or groups of loci are studied in pedigrees and linkage groups are
defined from the results. Two loci are said to be linked if the recombination frequency (R) is below 0.05.
Restriction fragment length polymorphisms (RFLPs) are useful markers for linkage studies. RFLP analyses rely on
the detection of alterations in the recognition sites of type II restriction endonucleases. If the RFLP which is linked
to the gene locus can be shown to be inherited concordantly with the disorder in a significant number of pedigrees,
it provides good evidence that the disorder is caused by alterations in the candidate gene. Problems can arise if
recombination occurs between the marker and the gene locus but this can be overcome by using a marker within
the candidate gene. Results can become confusing if more than one gene is involved in the pathology of the
disorder and linkage data become complex. This approach was used successfully to determine that osteogenesis
imperfecta is linked to the loci COL1A1 and COL1A2 (Sykes et al., 1986). Once a mutant locus has been
identified, it is possible to search for mutations underlying the disorders linked to the gene.
Mutation detection can be divided into two areas (i) the detection of unknown mutations and (ii) the
detection of mutations which have already been characterised *i.e.* for diagnostic purposes. Methods used for the
detection of mutations have been reviewed (Cotton, 1993; Grompe, 1993). Methods which are used for diagnostic
purposes are not discussed in this context. The detection of uncharacterised mutations can be further divided into
(a) methods which rely on the differences in electrophoretic properties between a normal and mutant sequence and
(b) methods which involve cleavage and chemical modification of heteroduplex DNA. The most commonly used
methods are those that are sensitive enough to detect single base changes as well as insertions and deletions,
therefore, methods such as *S*<sub>1</sub> nuclease mapping are not included in the discussion. The common methods in use
will be dealt with together in section 1.6.2. It is to be understood that these methods are designed to identify
changes in sequence away from the so-called normal sequence. Such a change may not be deleterious to the
individual — indeed some changes may be beneficial. Throughout the successive chapters of this thesis, sequence
changes which have resulted in a disease phenotype have been referred to as mutations. Those sequence changes
that alter a restriction site such that the frequency of the change warrants its classification as a restriction fragment
length polymorphism (RFLP), have been referred to as such. Sequence changes that occur in the third base position
of the codon and thereby do not alter the primary structure of the polypeptide are termed silent sequence variants.
Since most of the methods currently in use rely on DNA sequences amplified by the polymerase chain reaction, this
will be introduced first.

1.6.1 The Polymerase Chain Reaction (PCR)

The polymerase chain reaction (Salki et al., 1985) is a powerful method for the *in vitro* amplification of
DNA. Large amounts of a specific DNA fragment can be generated from a small amount of template. The method
requires a DNA polymerase and two oligonucleotide primers which define the outer limits of DNA sequence to be
synthesised. Usually, one of these primers is complementary to the sense DNA strand and the other complementary
to the antisense strand. A typical length for a primer is 20 nucleotides. The primers are designed such that their 3′
ends face the target sequence. It is not necessary to know the sequence of the target DNA as long as 5′ and 3′
flanking sequence is known in order to design suitable primers though a knowledge of the distance separating the
two priming sites can be useful. Some applications do not require knowledge of a specific sequence and rely on
degenerate (redundant) oligonucleotides based on consensus sequences or, amplification via vectorette (Arnold and
Hodgson, 1991) or 'panhandle' structures (Jones and Winstorfer, 1992), for example. The PCR is carried out in a
series of cycles; a denaturation step renders the DNA single stranded and a subsequent annealing step allows the
primers to anneal to their complementary target sequences. Finally, in an elongation step, each primer is extended
by the action of a DNA polymerase. These three steps are repeated a number of times until sufficient product has
been generated. Provided that each primer is capable of being extended through the sequence of the other one, the
product acts as a template for the next round of amplification. In this manner the amount of PCR product increases
in an exponential fashion as a function of the cycle number. After a few rounds of amplification, the major product
is equal to the length of DNA flanked by the oligonucleotides plus the sum of the lengths of the two primers.
Early amplification experiments (Saiki et al., 1985; Mullis and Faloona, 1987) utilised *Escherichia coli* DNA polymerase (the Klenow fragment) which required an annealing temperature of 37°C and thus needed to be renewed after every 95°C denaturation step due to its inactivation. The use of this enzyme also resulted in the generation of impure product due to the decreased primer specificity at the low annealing temperature. The advent of a heat-resistant DNA polymerase isolated from *Thermus aquaticus* (Taq DNA polymerase) allowed primer annealing to be carried out at much higher temperatures and it was not necessary to renew the enzyme at every cycle (Saiki et al., 1988) since the enzyme is still active after the many denaturation steps used. This has allowed the automation of the process and thermocyclers are now commercially available. There are problems with Taq DNA polymerase in that it lacks a 3’ to 5’ exonuclease activity which can cause the misincorporation of nucleotides due to lack of DNA synthesis proof-reading function (Tindall and Kunkel, 1988). If a base is misincorporated, both the incorrect and correct sequences will be amplified, which is unwanted in most applications except in forced mutagenesis (Gibbs, 1990). If such a mistake occurs early on in the process, the mutant sequence will be highly amplified, becomes a major reaction product and can be detected upon sequencing. Error rates have been estimated to be approximately 1:4000–1:5000bp of amplified, cloned and sequenced DNA (Innis et al., 1988) and as 2 x 10⁻⁴ incorrect bases added per nucleotide per amplification cycle (Saiki et al., 1988). This can be circumvented by the use of one of the alternative enzymes now available. Such enzymes include Vent® (Thermococcus litoralis), Vent® (exo'), Deep Vent® (Pyrococcus spesi er GB-D), Deep Vent® (exo') DNA polymerases. Deep Vent enzymes are more stable at high temperatures than Vent enzymes. The non-(exo') enzymes possess the exonuclease activity and thus reduce the number of misincorporated bases by approximately 27% (Cariello et al., 1991). The (exo-) enzymes lack this exonuclease activity and can be used for high temperature applications when a slightly higher base misincorporation rate is acceptable; this rate is still less than that of Taq DNA polymerase. The Deep Vent enzymes have half lives at 95°C and 100°C of approximately 3 and 8 times those of Vent enzymes. Other commercially available thermostable enzymes are Hot Tub, Tma, AmpliTaq, T7Th and AmpliTaq Exo Mutein polymerases. The properties of these enzymes have been reviewed by Ohler and Rose (1992). PCR can be used to amplify both genomic DNA and cDNA following a reverse transcription step (Veres et al., 1987). Taq DNA polymerase has been recognised to possess both reverse transcriptase and polymerase activities (Jones and Foulkes, 1989) and thus first strand cDNA synthesis and subsequent amplification can be effected by the single enzyme. This can be carried out at a higher temperature and can cause the disruption of secondary structure that may inhibit either reaction.

The PCR method itself can be used to screen for mutations via an application termed multiplex PCR. This involves the simultaneous amplification of several sequences throughout the gene of study and can be used to detect deletions which manifest themselves as the absence of expected fragments. This process has been used in the diagnosis of DMD-affected individuals in which approximately 60% of mutations are deletions. Heterozygous individuals are detected as a decrease in intensity of fragments compared to a normal individual.

Detailed information regarding the PCR has been collated (Erlich, 1989; Bloch, 1991; Amheim and Erlich, 1992).
1.6.2 Mutation detection by conformational changes

1.6.2.1 Single Strand Conformation Polymorphism (SSCP) analysis

This method was introduced by Orita et al. (1989a, b) and is widely used. It is based on the observation that single stranded DNA (ssDNA) in solution under certain conditions has a defined secondary structure which is dependent on its primary sequence; this structure may be altered when one of the bases is changed. The change in the secondary structure is detected by electrophoresis; under non-denaturing conditions the ssDNA has a mobility dependent on its sequence and its size and a mutant strand will have a different mobility to a wild type one. The main advantage is that the SSCP method is simple to use and is relatively sensitive (70–100% detection rate of mutations in PCR fragments below 350bp in length; Hayashi and Yandell, 1993). Many modifications to the original protocol have been introduced. Some researchers have eliminated the use of radioactivity in labelling the amplification products by the use of silver staining (Mohabeer et al., 1991; Ainsworth et al., 1991; Dockhorn-Dworniczak et al., 1991; Lenk et al., 1993) and ethidium bromide staining (Yap and McGee, 1992). Differential migration between two strands is most readily detected in fragments below 400bp (Hayashi, 1991) but greater than ~150bp (Sheffield et al., 1993) in length and the mutation can be identified by DNA sequencing. Strategies have been introduced such that larger fragments can be analysed (i) a larger fragment of DNA is cut with several restriction enzymes to produce fragments in the desired size range for analysis (Iwahana et al., 1992; Mackay et al., 1993a). This also serves to place the mutation in as many different sequence contexts as possible and thus may aid in its detection, (ii) a large fragment of DNA is cut with a frequent-cutting enzyme, separated according to size on a denaturing gel in the first dimension and then analysed on a non-denaturing gel in the second dimension by SSCP (Kovar et al., 1991). RNA has also been used as a substrate for SSCP (Danenberg et al., 1992; Sarkar et al., 1992a) since it is assumed that RNA can form a greater number of conformers which are sensitive to single base changes. Both reports on the RNA-SSCP technique indicate that it is more sensitive than DNA-SSCP. DNA-SSCP has been used to detect mutations in mildly affected cystic fibrosis individuals (Dean et al., 1990) and individuals suffering from non-insulin-dependent diabetes mellitus (O’Rahilly et al., 1991) as well as mutations in collagen genes (Mackay et al., 1993a; Rose et al., 1993).

1.6.2.2 Denaturing Gradient Gel Electrophoresis (DGGE)

When double stranded DNA is electrophoresed into an increasing gradient of denaturant (urea/formamide), a portion of the strand denatures but the strands are prevented from complete separation by a higher melting sequence-dependent domain(s) which does not melt at this concentration of denaturant. Once the duplex has partially unwound, its mobility is arrested in the gel. If a single base change exists in a similar duplex, the denaturation concentration for this duplex may differ in which case the duplex would migrate to a different position in the gel. This so-called band shift locates the presence of a mutation to a certain stretch of DNA (Myers et al., 1985a). If heteroduplexes between mutant and wild type molecules are used, the difference in mobilities is more noticeable. One modification made to the system was the inclusion of a 40 base pair GC-rich sequence (GC clamp) at one end of the fragment to be screened using a PCR-based approach (Myers et al., 1985b; Sheffield et al., 1989). Thus the whole area to be screened is a low melting point domain with respect to the clamp, which
increases the detection efficiency. The method is reported to be approximately 95% accurate in PCR products less than 600bp in length (Grompe, 1993) and as in SSCP analysis, the exact point of mutation has to be determined by sequence analysis. Many variations on the method have been documented (Cotton, 1993) including the generation of the denaturing gradient by temperature; temperature gradient gel electrophoresis (Tee et al., 1992) and temperature sweep gel electrophoresis (Yoshino et al., 1991).

1.6.2.3 Heteroduplex (HD) analysis

Heteroduplexes containing single base pair mismatches can be accurately separated from homoduplexes on non-denaturing polyacrylamide gels due to sequence dependent conformational changes in the dsDNA (Keen et al., 1991; Perry and Carrell, 1992; White et al., 1992). It is a simple method but since it is not 100% effective, Keen et al. (1991) suggested that Hydrolink™ gels should be used to maximise the mutation detection rate. It has been suggested that the sensitivity is approximately 80-90% for DNA fragments below 300bp in length (Perry and Carrell, 1992; White et al., 1992) and it is possible that SSCP and HD analyses could be used in tandem to try and bring the mutation detection rate closer to 100%. The method has been used successfully in identifying a mutation in Duchenne muscular dystrophy patients (Prior et al., 1993).

1.6.3 Mutation detection by heteroduplex mismatch

1.6.3.1 Ribonuclease (RNAase) cleavage

This method exploits the fact that many ribonucleases cleave single stranded RNA 3' to pyrimidine residues (Myers et al., 1985a). Single base pair mismatches in both RNA:RNA (Winter et al., 1985) and RNA:DNA (Myers et al., 1985c) heteroduplexes can be cleaved and the reaction products are analysed by electrophoresis. The presence and location of a mutation are indicated by a cleavage product band. The main advantage of the method is that it is a single step reaction, however a special RNA probe has to be made and the detection efficiency is only about 70%; this is because when purines appear in the probe at the mismatch, most mismatches fail to cleave. Due to these problems, the simpler chemical cleavage of mismatch method (CCM; section 1.6.3.2) has largely replaced its use. Despite the disadvantages, the method has been used successfully in identifying point mutations in ras genes (Winter et al., 1985) and the human immunodeficiency virus (Lopes-Salindez et al., 1991) as well as mutations in collagen genes (e.g. Hawkins et al., 1991; Molyneux et al., 1993; Marini et al., 1993a).

1.6.3.2 Chemical Cleavage of Mismatch (CCM)

This method developed by Cotton et al. (1988) relies on the formation of a heteroduplex between radiolabelled wild type DNA and mutant DNA or RNA. The mismatched or unmatched bases at the point of mutation are chemically modified. The method is based on the fact that mismatched C and T bases in heteroduplexes are more reactive with hydroxylamine and osmium tetroxide, respectively, than the equivalent matched base pairs which means that both heteroduplex species must be analysed in order to detect mismatched A and G bases. Radiolabelled DNA is cleaved by piperidine at the site of modification and cleavage products are
analysed by denaturing polyacrylamide gel electrophoresis. The method is very sensitive; more than 95% of mismatches can be detected if only the wild type DNA is labelled. This figure rises to 100% if both normal and mutant strands are labelled (Forrest et al., 1991). Silver staining has been used (Saleeba et al., 1992) in order to eliminate the use of radioactivity. Other major advantages are the relative lack of size constraints for template — PCR products of up to 1.7kb in length have been screened successfully (Zheng et al., 1991) — and the ability to determine the exact position of the mutation and the nature of the change by size of the cleavage product. However, the method requires many manipulations with mutagenic chemicals. It has been used for the detection of mutations causing disease e.g. p53 mutations in colorectal cancer (Rodrigues et al., 1990) and mutations in the COLIA1 gene (Valli et al., 1991; Bateman et al., 1993; Mottes et al., 1993), despite the disadvantages.

1.6.3.3 Carbodiimide modification (CDI)

A variation of the CCM method, carbodiimide modification, was proposed by Novak et al. (1986). Carbodiimide is a bulky reagent that interacts with the imino sites of T and G bases and it was shown that mismatched T and G bases in DNA heteroduplexes react more readily than the matched ones. Thus when carbodiimide is reacted with heteroduplexes between mutant and wild type sequences, any base change should potentially be detected by electrophoresis on a high percentage polyacrylamide gel. It has the advantage over CCM in that it is a single step reaction. An adaptation by Ganguly and Prockop (1990) employed the carbodiimide modification step but followed by primer directed extension using radiolabelled oligonucleotides in both directions. Taq DNA polymerase does not extend through modified bases and the shortened amplification product from the mutant allele is detected by electrophoresis and autoradiography and this approach was quoted as being able to localise the mutation to within 15bp. This adaptation of the method has been successfully applied to the detection of mutations in collagen genes (Zhuang et al., 1991; Ganguly et al., 1991; Spotila et al., 1991).

1.6.4 Mutation detection by sequencing

The use of manual sequencing for the screening of DNA for mutations would appear to be impractical. However, both alleles from 50 patients with aortic aneurysms were sequenced to identify the defect in the type III collagen gene (Tramp et al., 1993). In general though, the method is coupled with another mutation detection system which localises the mutation to a smaller region which is then subjected to sequence analysis. None of the methods described for the detection of mutations define the precise nature of the sequence defect. The definitive detection system for a mutation is the determination of the nucleic acid sequence. DNA sequencing can be carried out either by (i) cloning a fragment of interest into a vector and using universal or specific primers coupled with the chain termination reaction (Sanger et al., 1977), or (ii) directly sequencing a PCR amplified product (reviewed by Bevan et al., 1992). Different methods for the latter exist including one using the thermostable Taq DNA polymerase (Peterson, 1988; Innis et al., 1988). At present sequencing is a multi-step process although it is able to detect all mutations. However, with the advent of automated sequencers this may not remain such a problem.
1.7 AIMS OF THE STUDY

The aims of this study were to characterise mutations in the COL1A2 gene which could account for the OI phenotype in a number of affected individuals. In doing so it was hoped that information regarding such mutations would add substantially to the limited number identified in this gene and would complete a study of mutations in type I collagen genes initiated in the laboratory. In addition, a novel source of collagen gene transcripts was investigated with the view to using it (i) in the isolation of novel collagens using a PCR-generated collagen library approach and (ii) as a non-invasive tissue source for the determination of mutations causing OI and potentially other disorders.
CHAPTER TWO
MATERIALS AND METHODS

2.1 SOURCES OF MATERIALS

2.1.1 RNA, DNA and fibroblast cell lines

RNA and DNA from all cell lines from patients with various forms of osteogenesis imperfecta were kindly donated by Dr. K. Mackay (Mackay, 1992) with the exception of RNA from patient 'BF' which was kindly provided by Dr. Anne De Paepe (Faculty of Medical Genetics, University of Gent, Belgium). Fibroblast cell lines 89-247 and 89-248 were provided by Prof. Peter H. Byers (Departments of Pathology and Medicine, University of Washington, Seattle, USA) and that from patient 'PE' was supplied by Prof. Beat Steinmann (Department of Paediatrics, University of Zurich, Switzerland). Normal DNA control samples had been previously prepared from various anonymous donors from the Leicester Royal Infirmary, Leicester (placental DNA) and members of the Department of Genetics, University of Leicester, Leicester (blood DNA). Lymphoblast RNA was donated by Dr. K. Molyneux (Molyneux, 1991). Lymphoblastoid cell line DNA from family number 1362 was obtained from Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. MS32 probe DNA was obtained from Prof. Alec Jeffreys (Department of Genetics, University of Leicester, U.K.).

2.1.2 Synthetic oligonucleotides

Oligonucleotides used as primers in sequencing and the polymerase chain reaction were synthesised either by John Keyte (Queen's Medical Centre, Nottingham) or Debra Langton, (Department of Biochemistry, University of Leicester) using an Applied Biosystems 380B DNA synthesiser.

2.1.3 Bacterial strains and plasmids

Recombinant and non-recombinant plasmids were transformed into a recA strain of *Escherichia coli* JM83 (Vieira and Messing, 1982). This strain was created by Dr. M. Matfield (Matfield, 1983) and was obtained from Prof. W. Brammar (ICI Joint Laboratory, University of Leicester). The plasmid vectors pTZ19R (Pharmacia LKB Biotechnology, Uppala, Sweden) and pBluescript SK' (Stratagene Cloning Systems, La Jolla, California, USA; Short *et al.*, 1988) were used for cloning of DNA sequences. The plasmid vector pUC13 (Messing, 1983) was used as size marker for polyacrylamide gel electrophoresis upon digestion with restriction enzymes.

2.1.4 Reagents and other materials

Only where it was thought that a product from a particular manufacturer might be important for the application or that a product was sufficiently uncommon to warrant details of its origin, is it included in the following list. Agarose — SealCem, SeaPlaque — (FMC Bioproducts); Bovine α-c-casein (Merck/Darmstadt), Glass beads (No. 11; Jencons (Scientific) Ltd., Leighton Buzzard), Hybond-Nf nylon filters (Amersham International plc.), Hydrolink™ MDE™ gel (AT Biochem, Malvern, PA), Quick Draw blotting paper (Sigma), RNase/DNase-free BSA (Pharmacia LKB Biotechnology), RNasin (Promega Ltd., Southampton), Saran™ Barrier
Food Wrap (Dow Chemical Company), 325 mesh silica (Ceramic Developments (Midlands) Ltd., Corby), thermostable (Taq) DNA polymerase (Advanced Biotechnologies), T4 DNA ligase and tissue culture reagents — Dulbecco’s Modified Eagle Medium, Foetal calf serum (heat inactivated), 1 × Tris-HCl, phosphate-buffered saline — (Gibco-BRL, Life Technologies, Inc., Paisley, Scotland).

2.2 GENERAL METHODS

2.2.1 Centrifugation

For simplicity, centrifugation speeds are followed by an abbreviation, in parentheses, which indicates the type of centrifuge used according to the table below.

<table>
<thead>
<tr>
<th>CENTRIFUGE</th>
<th>ROTOR (IF APPLICABLE)</th>
<th>ABBREVIATION USED IN TEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorvall RT6000D</td>
<td>H1000B</td>
<td>RT</td>
</tr>
<tr>
<td>Sorvall RC-5B</td>
<td>GS3</td>
<td>GS3</td>
</tr>
<tr>
<td>Sorvall RC-5B</td>
<td>HB4</td>
<td>HB4</td>
</tr>
<tr>
<td>MSE Centaur</td>
<td>4-place swing-out</td>
<td>Centaur</td>
</tr>
<tr>
<td></td>
<td>(43124-126)</td>
<td></td>
</tr>
<tr>
<td>MSE Micro Centaur</td>
<td>Microfuge</td>
<td></td>
</tr>
</tbody>
</table>

2.2.2 Digestion of DNA with restriction enzymes

Reactions were incubated, unless stated otherwise, at 37°C in the presence of the manufacturer’s recommended buffer for 90 minutes for plasmid or PCR amplified material or 3–6 hours for genomic DNA. If required, bovine α-casein was added to a final concentration of 100 μg/ml prior to incubation (Dreyer and Schulte-Holthausen, 1991).

2.2.3 Purification of DNA from low gelling temperature agarose gel

This method was communicated by Chris Upton via the Bionet 'Methods and Reagents' bulletin board.

2.2.3.1 Preparation of glass milk slurry

To 80 ml distilled water were added 40 g silica 325 mesh. This was stirred for one hour and allowed to settle for 90 minutes. Supernatant was removed and pelleted at 3000 rpm (RT) for 10 minutes. The pellet was resuspended in 20 ml of distilled water and concentrated nitric acid was added to 50% (v/v). The mixture was heated to 95°C in a water bath and then allowed to cool. The glass was pelleted as before and washed five times with distilled water. The final pellet was stored as a 50% (v/v) slurry in distilled water and stored at 4°C.
2.2.3.2 Preparation of the NaI solution

To 100 ml of distilled water were added 90.8 g of NaI and 1.5 g of Na$_2$SO$_4$. These were mixed thoroughly and filtered through a Whatman no. 1 filter paper. A dialysis bag containing 0.5 g Na$_2$SO$_4$ was placed in the container to keep the solution saturated. The solution was stored foil wrapped at 4°C.

2.2.3.3 Purification protocol

This was adapted from the GENECLEAN II kit (BIO 101) Inc. protocol.

The desired fragment was excised from a 0.6% low gelling temperature (LGT) agarose gel (section 2.3.1). To the gel was added approximately 3 volumes of NaI solution; this was then incubated at 65°C until the agarose had melted. 5 μl of glass milk slurry were added to the solution and mixed by brief vortexing. This was then incubated on ice for 5 minutes, flicking the tube every 1–2 minutes to keep the glass milk suspended. The silica matrix was pelleted by centrifugation at 13000 rpm (microfuge) for 5 seconds. The supernate was removed, the pellet was washed three times in 200 μl NGEW wash (100 mM NaCl, 1 mM EDTA, 50% (v/v) ethanol, 10 mM Tris HCl; pH 7.5) and recentrifuged as above. The pellet was resuspended in 5 μl distilled water and the DNA eluted at 65°C for 2 minutes with gentle agitation followed by centrifugation for 30 seconds. Supernate was removed to a fresh tube and the elution step repeated. The supernates were combined and the pellet discarded.

2.2.4 Phenol/chloroform extraction and ethanol precipitation

The procedures were those described by Sambrook et al. (1989) unless stated otherwise.

2.2.5 Gel photography

Gels stained with ethidium bromide were photographed on a UV transilluminator (Chromato-vue, model TM-40, Ultra-Violet products Inc., San Gabriel, California, USA) using a Polaroid MP-4 Land Camera (Polaroid Corp., Cambridge, Massachusetts, USA) and Kodak T max 100 professional film. The film was developed using LX 24 X-ray developer, PX-40 X-ray liquid fixer and HDX-40 X-ray liquid hardener (all from Eastman Kodak Company, Rochester, New York, USA).

2.2.6 Autoradiography

A single sheet of Fuji RX X-ray film was placed over the filters or gel in a light-proof cassette. An intensifying screen was used if $^{32}$P was the radioisotope employed and omitted if $^{33}$P or $^{35}$S were the radioisotopes. If the screen was required, the cassette was placed at -80°C until the required signal intensity was achieved. If a sharper signal was required or $^{33}$P or $^{35}$S were the radioisotopes used, the cassette was left at room temperature.

2.2.7 Quantitation of nucleic acids

Small quantities of DNA or RNA were analysed by direct comparison under UV light with known amounts of DNA or RNA on an agarose gel following ethidium bromide staining. For increased accuracy, the absorbance of UV light was measured at 260 nm on a U-2000 Double-Beam Spectrophotometer (Hitachi Ltd.,
Tokyo, Japan). Nucleic acid concentration was determined using an O.D. of 1 unit corresponding to a concentration of 30 µg/ml for single-stranded DNA, 40 µg/ml for RNA and 50 µg/ml for double-stranded DNA.

2.3 GEL ELECTROPHORESIS

2.3.1 Routine agarose gel electrophoresis

Horizontal gel apparatus was used for agarose gel electrophoresis. SeaKem high gelling temperature (HGT) agarose was used to a maximum of 1.2% to separate DNA fragments of at least 100 bp in length. If the DNA was to be gel purified (section 2.2.3) a 0.6% SeaPlaque low gelling temperature (LGT) agarose gel was cast. An appropriate amount of agarose was dissolved in 1 x TAE buffer (40 mM Tris-acetate; pH 7.7, 1 mM EDTA) in a microwave oven. The agarose solution was cooled to approximately 55°C and EtBr was added to 0.2 µg/ml. It was then poured into the gel tray and allowed to set. Electrophoresis was carried out in 1 x TAE buffer at a constant voltage of between 1 and 12 V/cm gel length depending on separation and speed required. DNA markers used were λHindIII and φX174RF/HaeIII digest fragments. Samples were mixed with 0.2 volumes loading dye (20% (w/v) Ficoll 400; 0.1 mM EDTA, pH 8.0; 0.1% (w/v) orange G dye).

2.3.2 Non-denaturing polyacrylamide gel electrophoresis

For small scale analyses of DNA fragments which required superior separation to that achieved by agarose gel electrophoresis vertical polyacrylamide (PA) gels were used. The Bio-Rad Protean™ II and MiniProtein II™ tanks were used according to the manufacturer’s instructions. Depending on the degree of separation required, 7–10% gels were used. Gels were prepared by adding the appropriate volume of a 40% acrylamide stock (19:1 ratio acrylamide: bisacrylamide) to 0.5 x TBE buffer (90 mM Tris, 90 mM boric acid, 2 mM EDTA; pH 8.3). The acrylamide was polymerised by the addition of TEMED and APS to final concentrations of 0.032% (v/v) and 0.047% (w/v) respectively at room temperature. Samples contained 0.2 volumes loading dye (as above) and were electrophoresed alongside φX174RF/HaeIII or pUC13/HpaII digest fragment size markers at constant current and between 200 and 280 V. If the Protean II™ kit was used, cooling water was circulated through the apparatus. Following electrophoresis, gels were stained with ethidium bromide by agitation in 0.5 x TBE buffer containing 0.2 µg/ml EtBr for 30 minutes followed by washing in dH2O for 15 minutes. DNA was visualised by UV illumination.

2.4 TRANSFER OF DNA TO NYLON MEMBRANES

This was carried out using the method of Southern (1975) for agarose gels exceeding 16 cm x 12 cm in size or that cited by Lichtenstein et al. (1990) for gels smaller than this stated size except that Quick Draw blotting papers were used to effect transfer instead of paper towels. In both cases once transfer was complete, the apparatus was disassembled and the filter allowed to air dry. It was then wrapped in Saran wrap ensuring that the side with
the DNA on was covered by a single thickness only. The DNA was UV cross-linked onto the filter which was stored at 4°C.

2.5 HYBRIDISATION

2.5.1 Radiolabelling of DNA fragments

This method is based upon that of Feinberg and Vogelstein (1983).

The DNA to be labelled (8–10 ng) was made up to a volume of 8.8 µl with distilled water and placed in boiling water for 5–10 minutes. The labelling reaction was carried out by the addition of the following reagents in the given order at room temperature: 3 µl oligo labelling buffer (OLB), 0.6 µl 10 mg/ml DNase/RNase-free BSA, 2 µl [α-32P]dCTP (specific activity > 3000 Ci/mmol-1, 10 µCi/µl-1), and 0.6 U DNA polymerase I (Klenow fragment).

OLB was prepared from the following components: Solution A: 625 µl 2 M Tris-HCl; pH 8.0, 25 µl 5 M MgCl₂, 350 µl dH₂O, 18 µl β-mercaptoethanol, 5 µl each of 100 mM dATP, dGTP, dTTP, Solution B: 2 M HEPES; pH 6.6, and Solution C: hexadeoxyribonucleotides evenly suspended in 3 mM Tris-HCl, 0.2 mM EDTA; pH 7.0 at 90 O.D. units ml⁻¹. Solutions A, B and C were mixed together in the ratio of 2:5:3 respectively and stored at -20°C.

The labelling mixture was left at room temperature for a minimum of nine hours. The reaction was stopped by the addition of 85 µl of probe stop solution (20 mM NaCl, 20 mM Tris-HCl; pH 7.5, 2 mM EDTA, 0.25% SDS). Unincorporated nucleotides were not removed.

2.5.2 Checking the incorporation of a radio-labelled probe

From the stopped probe reaction 1 µl was removed to a fresh tube diluted with 9 µl of water and mixed thoroughly. Two pieces of Whatman DE-81 paper, 1.5 cm² in size, were labelled with the initials T (Total) and P (Precipitable); 1 µl of the diluted probe solution was pipetted onto each and both were allowed to air-dry. The piece labelled P was washed six times, five minutes per wash, with 0.5 M Na₂HPO₄. The ratio of counts (P:T), as measured by a Geiger counter, was estimated. The probe was not used if the ratio was below 40%.

2.5.3 Hybridisation of a radio-labelled probe to a nylon membrane

The hybridisation of a nylon membrane requires two steps; the pre-hybridisation step and the hybridisation. The pre-hybridisation step was different for non-genomic and genomic DNA blots.

2.5.4 Pre-hybridisation of a non-genomic DNA blot

The nylon filter was washed with a pre-wash solution containing 1.5 × SSPE (0.27 M NaCl, 15 mM NaH₂PO₄, 1.5 mM EDTA), 0.5% dried milk (Cadbury's Marvel), 1% SDS, 6% (w/v) polyethylene glycol 6000, at 65°C in a hybridisation oven for one hour.
2.5.5 Pre-hybridisation of a genomic DNA blot

The nylon filter was washed under the same conditions as for a non-genomic filter but in a pre-wash containing 1 mM EDTA, 0.5 M Na2HPO4; pH 7.2, 7% SDS. In both cases, the hybridisation solution was the same as the pre-hybridisation wash. The probe solution was denatured in boiling water for 5–10 minutes prior to adding to the hybridisation solution. Hybridisation was carried out at 65°C overnight in a shaking water bath.

2.5.6 Post-hybridisation washes

The filter was washed briefly three times in 3 x SSC (450 mM NaCl, 45 mM sodium citrate; pH 7.0), 0.1% SDS at 65°C. The filters were then washed twice for 10 minutes in the same manner. A more stringent set of washes (four times, 10 minutes each of 0.5 x SSC, 0.1% SDS) was used if necessary. The filters were allowed to air-dry.

2.5.7 Removal of radio-labelled probes from nylon filters

The probe was removed by washing the filters at 45°C for 30 minutes in 0.4 M NaOH and at 45°C for 30 minutes in 0.1 x SDS, 0.1 x SSC, 0.2 M Tris-HCl (mixed and adjusted to pH 7.5). The filters were allowed to air-dry and were autoradiographed overnight to check the efficiency of removal.

2.6 RADIO-LABELLING OF MARKER DNA FRAGMENTS

The bacterial plasmid pUC13 (500 ng) was digested with HpaII for one hour in a 20 μl reaction volume. To this digest were added 1 μl [α-32P]dCTP, 130 mM NaCl, 0.5 U Klenow and the volume adjusted to 30 μl. This mix was left at room temperature for 15 minutes then the DNA was ethanol precipitated and resuspended in 10 μl dH2O and 5 μl dye added (section 2.3.1).

2.7 CLONING OF DNA FRAGMENTS

If the fragment of interest was to be cloned into a single restriction site on a vector, the plasmid was phosphatased to prevent reannealing of the compatible vector ends.

2.7.1 Phosphatasing of the vector

The vector was digested with a suitable enzyme as in section 2.2.2, and the DNA purified as in section 2.2.3. The DNA was then incubated at 37°C for 30 minutes with 1 U calf alkaline phosphatase in the manufacturer's recommended buffer. The reaction was terminated by heating at 65°C for 10 minutes and the plasmid recovered as before.
2.7.2 Isolation and ligation of sequences of interest

DNA was digested with appropriate enzyme(s) and the fragment of interest purified as in section 2.2.3. In a final volume of 100 µl, 50 ng cut vector was mixed with 100 ng fragment of interest, and 1 U T4 DNA ligase in the manufacturer’s recommended buffer. This ligation mix and suitable controls were left at room temperature overnight.

2.7.3 Transformation of recombinant DNA

Two methods were employed:

2.7.3.1 Calcium chloride transformation of E. coli.

The following method, based upon that of Hutchison and Halvorson (1980), was used when a high transformation efficiency was not required.

Preparation of competent cells

A 3 ml volume of Luria broth (LB; 10 g Bacto tryptone, 5 g Bacto yeast extract, 5 g NaCl in 1 litre of water) was inoculated with a single colony of E. coli JM83 (recA) and was shaken for approximately 15 hours at 37°C. A 50 ml volume of fresh broth was inoculated with 1 ml of the small culture and was grown to an O.D. of 0.36–0.44 at 560 nm. The culture was centrifuged at 3000 rpm (RT) for 5 minutes at room temperature. The resulting pellet was resuspended in 20 ml ice-cold 0.05 M CaCl₂ and placed on ice for 15 minutes. The cells were recentrifuged and resuspended in 5 ml ice-cold 0.05 M CaCl₂, 5% glycerol. Competent cells were divided into 200 µl aliquots, snap frozen on dry ice and stored at -80°C until required.

Transformation procedure

Cells were thawed on ice and used immediately. Ligated DNA (approximately 10 ng) was resuspended in 10 µl freshly prepared transformation buffer (10 mM Tris-HCl; pH7.4, 10 mM MgCl₂, 10 mM CaCl₂). To this DNA were added 200 µl thawed cells and the mixture was incubated on ice for 25 minutes, heat-shocked at 37°C for 1.5 minutes and incubated at room temperature for 10 minutes. LB (1 ml) was added and following incubation with agitation for one hour at 37°C, the cells were plated onto Luria agar plates (LA; LB plus 15 g Bacto agar per litre) containing selective antibiotics and chromogenic indicator reagents (X-gal and IPTG) and grown overnight at 37°C.

2.7.3.2 Transformation of E. coli by electroporation

When a high transformation efficiency was required, electroporation was employed. The following method is based upon that of Dower et al. (1988).
Preparation of competent cells

A culture was inoculated as above and grown to an optical density of 0.55 at 560 nm. The cells were pelleted at 4000 rpm (RT) for 15 minutes at 4°C and resuspended in an equal volume of ice cold, sterile 10% glycerol. This was repeated three times, the cells first being resuspended in a half volume then in 1/50 volume and finally in 1/500 initial volume. The cells were snap frozen as above and stored at -80°C.

Transformation procedure

The DNA to be transformed was ethanol precipitated and resuspended in TE buffer (200 mM Tris, 10 mM EDTA; pH 7.6) or distilled water. The electroporator (BIO-RAD Gene Pulser™) was set up such that the capacitance was set at 25 μF and the volts at 1.5 kV. On ice, 1–5 μl transforming DNA was mixed with 40 μl washed cells. The mixture was transferred to a pre-cooled, washed cuvette and the cells tapped to the bottom of the cuvette. Once the pulse had been delivered to the cuvette, cells were allowed to recover by the addition of 1 ml SOC medium (2% w/v Bacto tryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCL, 10 mM MgCl₂, 10 mM MgSO₄, 20 mM glucose). The cuvette was rinsed thoroughly between successive pulses with distilled water to prevent cross-contamination of samples. The cells were slacked at 37°C for 1 hour and then plated onto LA plates containing selective antibiotics and chromogenic indicator reagents (X-gal and IPTG) and grown overnight at 37°C.

2.8 ISOLATION OF PLASMID DNA

2.8.1 Large scale plasmid preparation

This was carried out using a Qiagen plasmid kit (Qiagen GmbH, FRG) according to the manufacturer’s instructions.

2.8.2 Small scale plasmid preparation

One of two methods was used where appropriate. The method of Serghini et al. (1989) was used, without modification, for routine isolations. The second method is based upon that of Ish-Horowicz and Burke (1981) and was used when higher quality isolations were required. Bacteria were harvested as above, the supernate removed and the pellet gently resuspended in 100 μl solution I (50 mM glucose, 25 mM Tris-HCl; pH 8.0, 10 mM EDTA). This suspension was incubated at room temperature for 5 minutes after which, 200 μl freshly prepared denaturing solution (0.2 N NaOH, 1% SDS) were added, mixed gently and incubated on ice for 5 minutes. To this were added 150 μl pre-cooled 5 M potassium acetate; pH 4.8; the mix was vortexed briefly and placed on ice for 5 minutes. Following incubation, the mix was centrifuged at 13000 rpm (microfuge) and 400 μl of the supernate transferred to a fresh tube. RNaseA was added to a final concentration of 50 μg/ml and incubated at 37°C for 30 minutes. An equal volume of phenol/chloroform was added, vortexed and centrifuged as above for 2 minutes. The aqueous phase was transferred to a fresh tube and 2.5 volumes ice-cold 100% ethanol were added, mixed and incubated at -80°C for 30 minutes. Following incubation, the mix was centrifuged, as above, for 10 minutes and the
supernate removed. The pellet was washed with 70% ethanol and re-centrifuged for 1 minute, the supernate was discarded, the pellet vacuum dried and resuspended in 20 µl d.H₂O.

2.9 PLASMID LIBRARY SCREENING

This is based on a method cited by Buluwela et al. (1989).

Potential recombinant plasmids were analysed by transfer to a nylon membrane followed by hybridisation analysis. Single colonies were picked into separate wells in a microtitre plate containing 100 µl freezing medium (LB containing 0.1 volume HMFM — 3.6 mM K₂HPO₄, 1.3 mM KH₂PO₄, 2 mM trisodium citrate, 1 mM MgSO₄, 4.4% w/v glycerol). The plates were incubated at 37°C overnight in a sealed plastic bag containing a tissue soaked in water to prevent desiccation. Microtitre plates were stored at -80°C until required.

When the plates were needed they were allowed to thaw at room temperature. A piece of nylon membrane (Hybond-N+) was placed on an agar plate and using a multipronged tool ("hedgehog"), the colonies were transferred onto the filter in duplicate, the second application being offset with respect to the first. The hedgehog was dipped in IMS and flamed prior to each application. The colonies were allowed to dry after which the plates were incubated in an inverted position at 37°C overnight. The growth of the colonies was noted either as 'good', 'minimal' or 'none'. The filters were prepared for hybridisation analysis. They were laid sequentially, colony-side up, on pieces of filter paper impregnated with 10% SDS, denaturing solution (500 mM NaOH, 1.5 M NaCl), neutralising solution (3 M NaCl, 500 mM Tris-HCl; pH 7.4) and 3 x SSC for 3, 5, 5 and 5 minutes respectively. They were then allowed to air dry. The DNA was fixed onto the membranes by UV cross-linking of the filters and hybridised as in section 2.5.

2.10 NUCLEIC ACID PREPARATION FROM BLOOD

2.10.1 Preparation of genomic DNA

The method used is that described by Parzer and Mannhalter (1991).

2.10.2 Preparation of total mRNA

This method is based upon that of Chomczynski and Sacchi (1987).

In the preparation of RNA and subsequent applications involving RNA, it was necessary to reduce the amount of RNases in solutions and on equipment. In order to do this a DEPC solution was prepared by the addition of approximately 50 µl DEPC to one litre of d.H₂O. This was shaken vigorously and left overnight and the DEPC was degraded by autoclaving. Tubes were filled with the DEPC-treated water and tips were soaked in the solution overnight after which time the solution was discarded and the plasticware autoclaved. Solutions for use with RNA were prepared using DEPC-treated d.H₂O.

Blood was collected into citrate tubes and either used immediately or snap frozen and stored at -20°C. Frozen blood was rapidly thawed at 37°C when required and used immediately. The blood was mixed with
3 volumes of red cell lysis buffer (155 mM ammonium chloride, 10 mM potassium bicarbonate, 0.1 mM EDTA) and placed on ice for 10 minutes. The white blood cells were harvested by centrifugation at 800 rpm (RT) for 10 minutes. The pellet was resuspended in 1 ml denaturing solution (solution D; 4 M guanidinium thiocyanate, 25 mM sodium citrate; pH 7.0, 0.5% sodium sarcosyl, 0.1 M β-mercaptoethanol) and macerated using a Polytron (Kinematica GmbH, Switzerland) for 10 seconds. The mixture was transferred to a siliconised Corex centrifuge tube and to it were added 0.4 ml 2 M sodium acetate, 4 ml water-saturated phenol and 0.8 ml chloroform/IAA (49:1 mixture). This was shaken for 10 seconds and placed on ice for 15 minutes before being centrifuged at 9000 rpm for 20 minutes at 4°C (HB4). The aqueous phase was removed to a fresh siliconised Corex centrifuge tube, an equal volume of propan-2-ol was added, mixed and incubated at -20°C for at least 2 hours, preferably overnight. The RNA was precipitated by centrifugation at 9000 rpm for 20 minutes at 4°C (HB4). The pellet was resuspended in 300 µl solution D and was transferred to a DEPC-treated 1.5 ml microfuge tube. An equal volume of propan-2-ol was added prior to incubation at -20°C for 45 minutes. The RNA was pelleted at 13000 rpm for 20 minutes at 4°C (microfuge) and washed with 70% ethanol and dissolved in 40 µl DEPC-treated distilled water by pipetting and incubation at 65°C for 2–3 minutes. A fraction of the RNA was electrophoresed through a 0.8% agarose gel to determine the quality of the preparation. The RNA solution was stored at -80°C.

When RNA was subsequently used in experiments, care was taken not to introduce any RNase by using a clean pair of gloves and DEPC-treated dH₂O, tips and microfuge tubes.

2.11 AMPLIFICATION OF DNA SEQUENCES USING THE POLYMERASE CHAIN REACTION (PCR)

To eliminate the possibility of contaminating DNA in the PCR amplification process, care was taken to use clean reagents and equipment. Separate stocks of all solutions used in PCR reactions were maintained. Those that did not contain nucleotides as well as pipettes, tips and tubes were irradiated with UV radiation for 30 minutes periodically. All plasticware was used directly from the manufacturers’ bags. Each time a reaction, or reagents that were for use in reactions, were prepared, all work surfaces were swabbed with IMS and a fresh pair of surgical gloves were worn.

Oligonucleotides employed in the amplification process are identified as coding primers if their sequence is that of the sense template DNA strand or non-coding primers if the sequence is that of the antisense template DNA strand.

2.11.1 Purification of deprotected nucleotides

The method used was that of Sawadogo and van Dyke (1991).

2.11.2 Amplification of genomic DNA

The amplification procedure was adapted from Kogan et al. (1987). Approximately 20–50 ng total genomic DNA was amplified, typically in a 50 µl reaction volume containing 1 × PCR buffer (44.7 mM Tris-Cl; pH 8.8, 11.1 mM ammonium sulphate, 4.5 mM MgCl₂, 6.7 mM
β-mercaptoethanol, 4.5 mM EDTA; pH 8.0, 113 μg/ml RNase/DNase-free BSA), 1 mM each dATP, dCTP, dGTP and dTTP, an appropriate concentration of coding and non-coding primers (see individual applications) and 1 U Taq DNA polymerase. This mix was overlaid with paraffin oil to prevent evaporation. Amplification was carried out in a Programmable Dri-Block, PHC-3 (Techne Ltd.) for approximately 30 cycles consisting of a denaturation step at 95°C, an annealing step and an extension step at approximately 70°C. The time period for the extension step was dependent upon the length of template delimited by the coding and non-coding primers. The annealing temperature was dependent upon the dissociation temperature of the primers and was usually determined empirically. Primer details and amplification conditions for primer pairs are given at appropriate points in the text.

2.11.3 Amplification of RNA

RNA sequences were amplified via a cDNA intermediate. First-strand cDNA was synthesised essentially as described by Kawasaki and Wang (1989). A 20 μl reaction volume consisting of 1 x reverse transcription buffer (50 mM KCl, 20 mM Tris-HCl; pH 8.0, 2.5 mM MgCl₂, 100 μg/ml RNase/DNase-free BSA), 1 mM each dATP, dCTP, dGTP and dTTP, 20 U RNasin, 6 mM random hexadeoxyribonucleotide solution and 200 U M-MLV reverse transcriptase was prepared. This was mixed thoroughly and incubated at room temperature for 15 minutes, at 37°C for one hour and 95°C for 10 minutes. DNA was amplified in a 25 μl reaction volume consisting of 5 μl cDNA template, 1 x PCR buffer, 1 mM each dATP, dCTP, dGTP and dTTP, 0.72 μM each of coding and non-coding primer, and 0.5 U Taq DNA polymerase. This reaction was overlaid with paraffin. Conditions for amplification are again indicated at appropriate points in the text.

2.11.4 Reamplification of sequences

Reamplification of DNA was effected as above using approximately 5–10 ng original product to seed the reaction. The same amplification cycles were used as for the primary reaction but approximately half the number.

2.12 DNA SEQUENCING

2.12.1 Direct sequencing of PCR products

The chain termination procedure was used (Sanger et al., 1977) in a procedure adapted from that described by Innis et al. (1988). A 50 μl reaction volume was amplified as in 2.11.2, the product run on a LGT agarose gel and the correct fragment excised and purified as in 2.2.3. An asymmetric PCR (adapted from Gyllensten and Erlich, 1988) was carried out using one of the two primers employed in the initial reaction: to a 0.5 ml microcentrifuge tube were added 200–500 ng of cleaned PCR product, 1 μM primer 1, 1 x PCR buffer, 1 mM each dNTP and 2 U Taq DNA polymerase in a final volume of 50 μl. The DNA was amplified for 15 cycles using the same conditions as in the original amplification. The single-stranded DNA was precipitated by the addition of 1 volume 4 M ammonium acetate, 2 volumes propan-2-ol and incubation at room temperature for
10 minutes. The DNA was pelleted at 13000 rpm in a microcentrifuge, washed in 80% ethanol and vacuum dried. A 0.5 µl aliquot was electrophoresed through an agarose gel to determine the concentration.

The opposite primer to that used in the asymmetric PCR was radioactively end-labelled in the following reaction volume: 0.5 µM primer, 1 x kinase buffer (50 mM Tris pH 7.5, 10 mM MgCl2, 0.1 mM spermidine, 0.1 mM EDTA), 5 U T4 polynucleotide kinase and 1 µl [γ-^32P]ATP (specific activity > 3000 Ci mmol^-1); 10 µCi/µl) in a 20 µl reaction volume. This was incubated at 37°C for 1 hour.

Sequencing reactions were set up as follows: 1 x PCR termination mix (either A, C, G or T, see below), 1 µl DNA, 0.5 U Taq DNA polymerase and 1 µl end-labelled primer in a 7.5 µl reaction volume. These reactions underwent 10 cycles of amplification using the same conditions as in the original PCR reaction. The reactions were stored at -20°C if not required immediately. To 5 µl of each reaction was added an equal volume of dye (95% formamide, 20 mM EDTA, 0.05% (w/v) xylene cyanol, 0.05% (w/v) bromophenol blue) prior to denaturation at 80°C for two minutes and gel loading.

**Termination mix composition.**

The 20 x base mix consisted of 101.5 µl of 11.1 x PCR buffer (section 2.11.2) plus 1 µl each of 16 mM stocks of dATP, dCTP, dGTP and dTTP. Each 4 x termination mix consisted of 10 µl 20 x base mix and either 250 µM ddATP (A), 160 µM ddCTP (C), 80 µM ddGTP (G) or 330 µM ddTTP (T).

### 2.12.2 Sequencing of cloned sequences

Cloned DNA for sequencing was prepared as in section 2.7. Plasmid DNA was purified as in the second method of section 2.8.2 up until the end of the RNaseA incubation step. Following this the purified DNA was subjected to spin-dialysis as described by Murphy and Ward (1989) except that tubes were centrifuged at 1000 rpm (RT). DNA in solution was ethanol precipitated and resuspended in 10 µl dH2O.

The sequencing reactions were carried out as follows: to the denatured and neutralised DNA were added 1 x annealing buffer (50 mM Tris-HCl; pH 7.5, 20 mM MgCl2, 70 mM NaCl) and 5-10 ng sequencing primer. Following incubation at 37°C for 20 minutes, 1 x labelling mix (0.1 µM each dCTP, dGTP, dTTP), 10 mM DTT, 3 U T7 DNA polymerase and 0.5 µl [α-^35S]dATP (specific activity > 400 Ci mmol^-1, 10 µCi/µl) were added, mixed and the solution incubated at room temperature for 5 minutes. To four tubes, labelled ‘G’, ‘A’, ‘T’, ‘C’ were added 2.5 µl of the appropriate termination mix (50 mM NaCl, 80 µM each of dATP, dCTP, dGTP and dTTP and either 8 µM ddATP, ddCTP, ddGTP or ddTTP respectively) and pre-warmed at 37°C. To each of these tubes, 4.5 µl of labelling reaction were added. Following incubation at 37°C for 5 minutes 0.5 U E. coli DNA polymerase I (Klenow fragment) was added. After 5 minutes at 37°C, 5 µl stop solution were added and the samples were stored at -20°C unless required immediately.

### 2.12.3 Preparation of a sequencing gel

Sequencing was carried out using BRL Life Technologies Sequencing apparatus model S2 with 0.4 mm spacers and sharkstooth combs. One side of the small plate was siliconised and then both plates scratched with
detergent, washed thoroughly and wiped with IMS. The two plates were taped together with the spacers in place.

Sequencing gels were 6% acrylamide (from a 40% stock), 1× TBE buffer and 50% (w/v) urea. The urea was dissolved in the mixture at 65°C then filtered under vacuum through two Whatman no. 1 filter papers and allowed to cool to room temperature. Polymerisation of the acrylamide was initiated (section 2.3.2) and the gel poured into the mould. The combs were put in flat side down and the gel allowed to polymerise. The combs were removed, washed and replaced with the sharksteeth sunk into the gel to a depth of approximately 1 mm to form “wells”. The gel was then put in the sequencing tank and pre-run at 70–80 W for 30–60 minutes in 1× TBE. The wells were flushed out and 3 μl volume samples were heat-denatured at 80°C for 2 minutes and immediately loaded. The gel was run for a suitable length of time, judged by dye migration, at 70–80 W to maintain a running temperature above 50°C. After this time, the smaller plate was removed and the gel fixed in a solution of 10% methanol, 10% glacial acetic acid. The gel was transferred to a sheet of Whatman 3MM filter paper, dried under vacuum at 80°C and autoradiographed (section 2.2.6).

2.13 CULTURE OF FIBROBLAST CELLS

Non-transformed fibroblasts cultured from patient and parental skin biopsies were grown as a source of genomic DNA. Manipulation of cells was carried out in a class II microbiological safety cabinet with care being taken to avoid fungal and bacterial contamination or cross-contamination between cell-lines. All plasticware was used straight from the manufacturer’s packaging. Foetal calf serum and trypsin-EDTA were aliquoted into 20 ml sterile tubes and sterility tested before storage at -20°C. 1× phosphate-buffered saline (PBS) solution was diluted from the stock solution with sterile dH2O in the safety cabinet and filtered through a 0.2 μm disposable filter prior to use.

2.13.1 Growth and splitting of cultured fibroblast cells

Cells were grown as monolayers in 8 ml Dulbecco’s Modified Eagle Medium (DMEM) containing 25 mM HEPES buffer, sodium pyruvate plus 1 g l-1 glucose supplemented with 10% foetal calf serum (FCS) in 25 cm2 tissue culture flasks. The cells were incubated in a humid atmosphere at 37°C with 5% CO2; the lids were loosened to allow air exchange. The growth of cells was observed using a microscope at 40x under light or phase-contrast conditions. Confluent growth was said to have been reached when there was no space between neighbouring cells. For slow growing cell lines, the medium was replaced every 3–4 days. Cells were grown to confluency and then split at a ratio of 1:2. The medium was removed from a flask and the cells were washed twice with 2 ml of 1× PBS. This was removed and 1.5 ml trypsin-EDTA was added and the cells were incubated at 37°C until the cells had rounded up and just come away from the flask surface. A 16 ml volume of pre-warmed medium was added and 8 ml was removed to each of two clean flasks which were incubated as above.
2.13.2 Freezing and thawing cells

Cells were grown to subconfluency and were treated as for splitting except that trypsinisation was allowed to proceed only until the cells had rounded up but were still attached to the flask when the trypsin-EDTA was removed. Then, in quick succession, a 2 ml volume of medium containing 7.5% dimethylsulphoxide was added and the flask shaken to remove the cells from the bottom. Aliquots of 1 ml were transferred to cryotubes and placed next to a 500 ml beaker full of liquid nitrogen in a polystyrene box with a tight-fitting lid. After 20 minutes, the tubes were transferred to a dewar flask of liquid nitrogen for storage. When required, frozen cells were defrosted as quickly as possible in 37°C water with agitation and transferred, dropwise, into 7 ml pre-warmed medium supplemented with 20% FCS in a flask and incubated.

2.13.3 Preparation of genomic DNA from cultured fibroblasts

Cells were grown to confluency and treated as for splitting up to the point of trypsinisation. To the cells were added 3 ml medium and the mixture was centrifuged at 1100 rpm (Centaur) and the pellet resuspended in 1.4 ml 1 x SSC. Following centrifugation at 6500 rpm (microfuge) for 2 minutes, the cells were suspended in 200 ml suspension buffer (300 mM sodium acetate, 20 mM Tris-HCl; pH 7.5, 1 mM EDTA) and were lysed by the addition of 0.75% (w/v) SDS. The suspension was mixed gently for 2 minutes, added to 200 ml TE and phenol-chloroform extracted. The DNA was ethanol precipitated and resuspended in dH₂O.

2.14 SSCP ANALYSIS

This adaptation of the original SSCP method (Orita et al. 1989a, b) for use with collagen genes is essentially that described by Mackay et al. (1993a). However, samples were isopropanol precipitated by adding two volumes propan-2-ol and one volume 4 M sodium acetate. This was mixed and left at room temperature for 10 minutes before centrifuging for 10 minutes. The pellet was allowed to dry and was resuspended in 10 μl dH₂O. Samples for analysis were electrophoresed through a Hydrolink MDE gel containing 0.6 x TBE buffer, 10% glycerol at room temperature at 4 W. The running buffer was the same as that in the gel.

2.15 MINISATELLITE VARIANT REPEAT PCR

The method is essentially that of Jeffreys et al. (1991). Genomic DNA was typed for a- and t-type repeats of the minisatellite MS32. The amplification conditions differed in that 10 ng of DNA were amplified for 21 cycles using the given conditions. Autoradiography was at room temperature overnight.

2.16 COMPUTER PROGRAMS

DNA sequence analysis was carried out on an IRIX Challenger XL on the UNIX system Version 5.0.1 (Silicon Graphics International) using the Genetics Computer Group sequence analysis software package.
(Devereux et al., 1984). Design of PCR primers was assisted by use of the OLIGO package (Rychlik and Rhoads, 1989) on an IBM PC running MS DOS.
CHAPTER THREE

BACKGROUND TO THE STRATEGY FOR SCREENING FOR MUTATIONS IN PATIENTS WITH OSTEOGENESIS IMPERFECTA

3.1 INTRODUCTION

Single strand conformation polymorphism (SSCP) analysis has been widely used since its introduction by Orita et al. (1989a, b) and has been discussed in Chapter One. Many of its applications involve the electrophoresis of single PCR amplification products within a suitable size range for the detection of bands with altered mobility. This optimal size range has been suggested to be between approximately 100bp and 400bp (Hayashi, 1991; Sheffield et al., 1993). However, Iwahana et al. (1992) and Mackay et al. (1993a) generated fragments of a suitable length by endonuclease cleavage of a larger amplification product. The aim of this study is to identify mutations within human type 1 collagen genes resulting in an OI phenotype using the SSCP technique. These genes have a sequence in excess of 4kb encoding each of the two α-chains. It would be costly to synthesise oligonucleotide primers along the entire length of the mRNA such that amplification products of an optimal size for SSCP analysis were generated. It would be desirable instead, to use a limited number of primer pairs and be able to generate fragments of a suitable length for analysis from each one. The method of Iwahana et al. (1991) was expanded upon and adapted for use with collagen genes and had been shown to be successful for detecting mutations in the COL1A1 gene (Mackay, 1992; Mackay et al., 1993a). Thus it seemed appropriate to expand its use and apply it to the α2(1) collagen mRNA that, once all patients in the study had been screened, would enable a calculation of the efficiency of the method. Since the majority of mutations causing OI appear to be substitutions of vital glycine residues in the triple helical domain (section 1.5.1.1(a)) the analysis was focused on this region. In addition, it was expected that potential small deletions and insertions as well as exon skipping events within this region might be detected.

3.2 DETAILS OF PATIENTS USED IN THE SCREENING STUDY

Cell lines from 27 patients diagnosed by various collaborators as having an OI phenotype were available for investigation. All of these had been screened for the presence of a mutation in the COL1A1 gene which might have resulted in their phenotype (Mackay, 1992; Mackay et al., 1994). Of these, mutations in COL1A1 were characterised in nine patients and a single patient was screened for a mutation in the COL1A2 gene and was found to harbour such a mutation. The remaining 17 patients were screened for mutations in the COL1A2 gene as part of this study.

Determination of the extent of increased post-translational modification has been a helpful aid to determining the location of the mutation although not in identifying the gene in which the mutation occurred. This is because all chains in molecules that contain even a single abnormal chain are overmodified (if at all) N-terminal
to the abnormal sequence, reflecting the polarity of triple helix propagation. Cyanogen bromide peptide data for some of the patient cell lines were supplied by our collaborators. However, the search for mutations in these patients was not limited to the region predicted by peptide mapping to harbour the defect. This was so that an estimate of the efficiency of the SSCP analysis technique as applied to OI mutation detection could be made once all the patients had been screened for both type I collagen genes. Since the efficiency would be dependent on the sequence changes being in agreement with the peptide data, the whole of the triple helical region was screened for each patient. It is also likely that mutations in the C-terminus might be the cause of the OI phenotype. The severity of the phenotype will be determined by whether or not its chains are capable of interacting with other chains and are incorporated into a procollagen molecule. In such cases, CB peptide analysis would be expected to indicate overmodification of all peptide fragments. If the mutation occurred in the N-terminus, such that chain association and processing was unaffected, CB peptide data would reveal nothing as to the whereabouts of the mutation since it would be too N-terminal to cause any overmodification.

One more patient (BF) was added to the study. It was suggested by the collaborating laboratory that the patient's α1(I) CB peptide fragments were normal following two-dimensional electrophoresis but there was a tilted aspect to the α2(I) CB4 fragment. On this basis, the patient was screened only for mutations in the α2(I) mRNA in this study. Patients CF-2 to CF-4 are the affected sons of CF-1, the unaffected father. Since it is likely that the recurrence of the phenotype is due to germline or low level somatic mosaicism for a mutation, all three offspring will harbour the same mutation as the father or mother. Thus only patient CF-2 was used in the screening procedure. Only a paternal sample was supplied since, for obvious reasons, maternal germ cells are difficult to obtain. If a mutation was found all family members could be tested. CB peptide fragments for both type I collagen α-chains and their relation to the triple helical domain of the α2(I) mRNA is illustrated in figure 3.1. Details of all the patients and CB peptide data, where available, are presented in table 3.1.

3.3 THE SSCP STRATEGY

3.3.1 Localisation of the mutation

The adopted strategy involved the use of four overlapping primer pairs along the length of the α1(I) mRNA (Mackay et al., 1993a) generating amplification products of approximately 11b in each case. Sarkar et al. (1992a) proposed that the mobility of a single strand conformer derived from a mutant PCR product depends not only on the size of the fragment but also on the position of the defect within the fragment and the sequence composition of the DNA; a greater concentration of purines in the sequence may cause the fragment to migrate faster in a gel and show a more dramatic band shift (Sheffield et al., 1993). Thus, each amplification product was digested into a series of smaller fragments in single or double restriction endonuclease digest reactions. The strategy used for the α1(I) mRNA was adapted for the detection of mutations in the α2(I) mRNA (according to the sequence reported by Kuivaniemi et al., 1988b). This is the subject of the study discussed in Chapters Four to Eight. Oligonucleotide primers for use in the polymerase chain reaction were synthesised around existing primers in the α2(I) mRNA in order to generate four overlapping pairs spanning the triple helical region of the mRNA. The
Figure 3.1 Positions of \( \alpha 2(i) \) amplified fragments with respect to the triple helical peptide fragments

The upper two lines represent the peptide fragments derived from the CNBr digestion of the \( \alpha 1 \) and \( \alpha 2 \) chains of type I procollagen. The lower line indicates the relative positions of the \( \alpha 2(i) \) amplified fragments to the triple helical peptide fragments. STH, start of triple helical domain; ETH, end of triple helical domain.
### Table 3.1 Patient details

<table>
<thead>
<tr>
<th>PATIENT IDENTIFICATION</th>
<th>OI PHENOTYPE</th>
<th>CB PEPTIDE DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM02328 (23)</td>
<td>II</td>
<td>α1(I)CB8</td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td>II/III</td>
<td>α2(I)CB4</td>
</tr>
<tr>
<td>PI</td>
<td>III</td>
<td>α1(I)CB7</td>
</tr>
<tr>
<td>PIR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>III/IV</td>
<td>α1(I)CB7</td>
</tr>
<tr>
<td>SC</td>
<td>IV</td>
<td>α1(I)CB8</td>
</tr>
<tr>
<td>CF-2 GM02573 (25)</td>
<td>NON-LETHAL</td>
<td></td>
</tr>
<tr>
<td>GM01093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM02695 (95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM02645 (45)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The patient cell lines studied are given in the table. The patient phenotype is indicated where known. Where 'non-lethal' is cited, clinical data were not provided and thus a more precise classification was not possible. The third column indicates the most C-terminal α1(I)/α2(I)CB peptide fragment found to be overmodified (data supplied by collaborating laboratories). Where a patient is referred to by a truncated code in the text, the common identification is indicated in parentheses.
fragments resulting from such amplifications were identified as α, β, γ and δ from the 5' end of the coding strand.

The identification of the fragments and their positions in relation to the mRNA and CB peptide fragments of both type I collagen α-chains are illustrated in figure 3.1. For each patient, first strand cDNA was synthesised from the mRNA using random hexamers. From the cDNA template, each of the primer pairs was used to generate an amplification product of approximately 1kb; amplification conditions of the PCR and resulting fragment sizes are given in table 3.2. In each case, an informative restriction enzyme digest was performed on the fragment to ensure that the correct sequence had been amplified. If the correct fragment had been generated, an identical reaction was set up, this time incorporating [α-32P]dCTP. For each of the amplification products, a series of four single or double digests was carried out to generate a number of different fragments of an optimal size for SSCP analysis. The reason for carrying out multiple digests was that a mutation might not be detected as a band shift in any of the fragments resulting from a single digestion due to the nature of the "conformer" produced by the fragment (Mackay et al., 1993a) and thus multiple digests serve to place the mutation in a series of different sequence contexts which might aid its detection. The choice of restriction enzymes was made according to the predicted size of the resulting fragments and the degree of separation that would be achieved. Sizes of between 100bp and 400bp were aimed for although in some instances the inclusion of larger and smaller fragments was unavoidable. This was not thought to be problematic since if the mutation resides in a fragment that is not of an optimal size for the detection of a band with altered mobility, another digestion might place the mutation in a more suitably sized fragment for analysis. Fragments generated from each of the digests for all four amplification products are given in table 3.3.

In order to ascertain if the radioactive amplification reactions had been successful and as a check of the amplification fidelity, approximately 0.2 μl of each sample was electrophoresed through a 3 mm thick 0.8% agarose gel cast on a glass plate which was dried under vacuum and autoradiographed as in section 2.2.6. The results indicated if all the reactions had been successful and provided an estimate of the efficiency of the amplification as indicated by the signal intensity. If the efficiency of a given reaction appeared to be low relative to that of the other reactions, a larger amount of product was used in the digest reaction. If it displayed extremely low radioactivity incorporation (efficiency), the PCR was repeated. Radioactive digests were also verified by electrophoresis through an 8% mini acrylamide gel (section 2.3.2) to ensure that all the digests had gone to completion. The gels were treated as above. Any partial digests were discarded and the reaction repeated prior to SSCP analysis.

Digestion products were electrophoresed as in section 2.14. Hydrolink MDE™ (mutation detection enhancement) gel is a modified polyacrylamide-based vinyl polymer reputed to be a superior matrix for electrophoresis (Keen et al., 1991). It was observed that for a patient sample (PE) which was screened for defects in the α2(I) chain, a given band shift was detected in two out of a possible four digests when the samples were electrophoresed through a normal polyacrylamide SSCP gel, but in all four digests when the samples were electrophoresed through a Hydrolink MDE SSCP gel as outlined in section 2.14 (results not shown). In other samples (PJ, KO), the mutation manifested itself as two band shifts on normal polyacrylamide gels and three on MDE gels (results not shown). Hence it appeared that the MDE gel was more sensitive at detecting bands with altered mobility. The pattern of denatured (single stranded) fragments was more spread out than that seen on a
<table>
<thead>
<tr>
<th>PRIMER IDENTIFICATION</th>
<th>α2(1) SSCP FRAGMENT AMPLIFIED</th>
<th>SIZE OF PRODUCT (bp)</th>
<th>PRIMER SEQUENCE 5'→3'</th>
<th>POSITION OF PRIMER</th>
<th>AMPLIFICATION CONDITIONS</th>
<th>NUMBER OF CYCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSCP1A2αL</td>
<td>α</td>
<td>946</td>
<td>ATGGTGAGAGTGCTCCCACA</td>
<td>Exon 5</td>
<td>95°C – 1.4 min</td>
<td>35</td>
</tr>
<tr>
<td>A2B^1</td>
<td></td>
<td></td>
<td>TCACCGGTTGAAATCCGCTCA</td>
<td>Exon 21</td>
<td>63°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>...CCCTTGGTCGGC</td>
<td></td>
<td>70°C – 1.4 min</td>
<td></td>
</tr>
<tr>
<td>A2A^1</td>
<td>β</td>
<td>1039</td>
<td>TCCCTGGACCCCGGCCTT</td>
<td>Exon 19</td>
<td>95°C – 1.4 min</td>
<td>35</td>
</tr>
<tr>
<td>SSCP1A2βR</td>
<td></td>
<td></td>
<td>GCGTTACCAATTTCACCTCTG</td>
<td>Exon 33</td>
<td>55°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.4 min</td>
<td></td>
</tr>
<tr>
<td>SSCP1A2γL^2</td>
<td>γ</td>
<td>1141</td>
<td>GCCGCTCTACGCTAT</td>
<td>Exon 31</td>
<td>95°C – 1.4 min</td>
<td>34</td>
</tr>
<tr>
<td>SSCP1A2γR^2</td>
<td></td>
<td></td>
<td>GACCCGGTCTCAGGTGT</td>
<td>Exon 44</td>
<td>69°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.4 min</td>
<td></td>
</tr>
<tr>
<td>SSCP1A2δL</td>
<td>δ</td>
<td>700</td>
<td>CTTGCCATTTGGGCCCTTC</td>
<td>Exon 42</td>
<td>95°C – 1.4 min</td>
<td>35</td>
</tr>
<tr>
<td>SSCP1A2δR</td>
<td></td>
<td></td>
<td>CATAGTCTTGGTGCTGAGA</td>
<td>Exon 49</td>
<td>63°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.0 min</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2 Primers used for the SSCP analysis of the α2(1) collagen mRNA

Details of the oligonucleotides used to generate the four overlapping fragments along the length of the α2(1) collagen mRNA for SSCP analysis are presented. Conditions for amplification are given and may differ from those quoted in the original applications. ^1 Molyneux et al. (1993), ^2 Raghunath, Mackay, Dalgleish, Steinmann (manuscript in preparation).
### Table 3.3 Restriction digests of the α2(I) collagen cDNA

Details of restriction digests of all four α2(I) collagen mRNA SSCP fragments and the fragments generated are presented. The fragments are listed in order of descending size. Those marked with an asterisk have the 13bp tag at their 3' end derived from primer A2B.

<table>
<thead>
<tr>
<th>SSCP FRAGMENT</th>
<th>ENZYMES EMPLOYED</th>
<th>FRAGMENTS GENERATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>AluI</td>
<td>288, 246, 156, 114, 90, 52*</td>
</tr>
<tr>
<td></td>
<td>PvuII</td>
<td>360, 288, 156, 90, 52*</td>
</tr>
<tr>
<td></td>
<td>BglII</td>
<td>351, 296, 129*, 108, 72</td>
</tr>
<tr>
<td></td>
<td>NciI/PvuII</td>
<td>253, 156, 126, 116, 90, 89, 52*, 46, 18</td>
</tr>
<tr>
<td>β</td>
<td>HincIUSF1</td>
<td>247, 194, 189, 185, 137, 87</td>
</tr>
<tr>
<td></td>
<td>HincI/Ddel</td>
<td>239, 225, 189, 111, 97, 91, 65, 22</td>
</tr>
<tr>
<td></td>
<td>HinfI/MspI</td>
<td>321, 189, 120 (x 2), 87, 81 (x 2), 40</td>
</tr>
<tr>
<td></td>
<td>AluV/Ddel</td>
<td>304, 207, 142, 108, 91, 84, 36, 33, 22, 12</td>
</tr>
<tr>
<td>γ</td>
<td>MspI</td>
<td>387, 277, 236, 126, 113, 2</td>
</tr>
<tr>
<td></td>
<td>Ddel/SF1</td>
<td>350, 253, 235, 187, 116</td>
</tr>
<tr>
<td></td>
<td>AluIRsal</td>
<td>289, 266, 211, 195, 168, 12</td>
</tr>
<tr>
<td></td>
<td>Ddel/HincII</td>
<td>351, 216, 187, 165, 117, 70, 37</td>
</tr>
<tr>
<td>δ</td>
<td>MspII/Rsal</td>
<td>207, 184, 172, 126, 11</td>
</tr>
<tr>
<td></td>
<td>StyII/PstI</td>
<td>231, 192, 122, 101, 43</td>
</tr>
<tr>
<td></td>
<td>HincII/Rsal</td>
<td>395, 172, 70, 63</td>
</tr>
<tr>
<td></td>
<td>StyII/MspI</td>
<td>207, 126, 122, 101, 79, 43</td>
</tr>
</tbody>
</table>
normal acrylamide SSCP gel which could partly explain the increased ease of detecting shifted fragments with respect to the normal ones. On this basis, Hydrolink MDE gel was the chosen gel matrix for the analysis of all the samples. At present the nature of the mobility shift induced by a mutation can not be predicted but can vary with changes in electrophoretic conditions such as temperature, thus thermal conditions for all SSCP gels were kept constant. Since the separation achieved on an SSCP gel was expected to be much greater than that seen on the mini "check" gels, corresponding native and denatured samples were electrophoresed side by side. It was assumed that this would allow the identification of spurious PCR products as well as small deletions and insertions not visualised on the smaller gel. Native samples were not heated to 95°C prior to loading.

The results of the analyses carried out are presented in the following five chapters. In each case, when a band with altered mobility was detected on an SSCP gel, a fresh cDNA sample was reverse transcribed from total patient fibroblast RNA and the analysis repeated. This ensured that the results were reproducible and not artefacts. Gel artefacts could arise as a result of Taq DNA polymerase infidelity. The problems associated with this have been reviewed by Eckert and Kunkel (1991) and in Chapter One. If an error was introduced by the enzyme at an early stage of the PCR amplification it would be reproduced in an exponential fashion along with the correct sequences. This aberrant product would manifest itself as a strong signal on the SSCP gel and due to the sequence difference might be revealed as a band with altered mobility. If the experiment was repeated it would be highly unlikely that the same error would be made. In order to decrease the possibility of misincorporation the concentration of each of the dNTPs was decreased to one tenth that used in a regular PCR. This also elevated the degree of (α-32P)dCTP incorporation (Mackay et al., 1993a). All of the nucleotide concentrations were lowered as opposed to just dCTP since it has been observed that when one nucleotide is in excess over the others, the rate of base misincorporation may be increased (Eckert and Kunkel, 1991). Multiple artefacts were observed on early SSCP gels. It was seen that in some samples, for a given fragment from the restriction digest, more than one band was present in both the native and the denatured samples. The bands appeared to differ by ± 1 or 2bp from the expected fragment. This phenomenon was not witnessed in samples digested with enzymes which generate a blunt ended fragment and seen in fewer fragments derived from double digests which involved one enzyme which generated a blunt end and one which generated an overhanging end. Thus it was speculated that during the ethanol precipitation of radio-labelled amplification product a small amount of Taq DNA polymerase was carried over with the DNA. The 5' to 3' exonuclease activity of this enzyme meant that products with overhanging ends resulting from enzyme digestion were partially degraded — the Taq DNA polymerase removing the overhanging base pairs in various combinations from each end of the fragment — giving rise to fragments of differing lengths. To eliminate this problem, the amplified products were precipitated by the addition of isopropanol. This reduced the amount of Taq DNA polymerase carried over with the DNA and the extent of the degradation was seen to be considerably reduced. Although digested products were electrophoresed through a check gel prior to an SSCP gel, this degradation had not been observed. In retrospect this was not unexpected since the degree of separation of the fragments on an SSCP gel is far greater than that achieved on a small acrylamide gel.

If the second SSCP gel reproduced the results observed on the initial one, the data were collated and analysed. It was possible to predict which double stranded fragment had given rise to which single stranded
fragments. This was easiest when the digestion products were of varied size and when the single stranded fragments were widely separated. The assignment of bands from MDE SSCP gels was often easier than those on a normal polyacrylamide gel due to the fragments having even greater differing mobilities. It was a possibility that some bands could not be assigned with confidence to any one double stranded fragment due to them having similar sizes, however, this was not a problem encountered in this study due to the position of the sequence changes. In some cases, the mutation may not be evident as a distinct band shift but rather as a smeared product. This is possibly due to the single stranded DNA having multiple conformations which are in equilibrium; this could be expected to manifest itself as a widened band on the gel. Some shifts could be seen in both single stranded fragments from one double stranded fragment suggesting that the two single strands may each exist in two non-transitional conformations under the electrophoretic conditions of the gel.

When band shifts were observed, they could be assigned to a single stranded and hence a double stranded fragment. This information was transferred onto a restriction map of the amplified region. Assuming that more than one band with altered mobility had been detected, a region of overlap was identified and thus the area in which the mutation was expected to lie was predicted. Subsequent sequence analysis could then be confined to a smaller region.

### 3.3.2 Identification of the sequence defect

Regions containing potential mutations were sequenced either by cloning the region of interest into a bacterial plasmid or by performing Taq DNA polymerase cycle sequencing if the region was not amenable to cloning. If a sequence was determined from cloned products, several clones were analysed. Since the probands are expected to be heterozygous and individual clones were derived from total PCR product, the clones would be expected to have either a normal or a mutant sequence insert. Therefore a set of clones (n ≥ 8) was analysed each time. If more than one clone showed a certain mutant sequence the defect was assumed to be genuine. If only one clone displayed an altered sequence it was assumed to be a highly amplified PCR artefact. There were always expected to be clones containing the normal sequence therefore the simultaneous analysis of a known control individual was not necessary. If a sequence was determined by the direct analysis of the double stranded PCR product (cycle sequencing) only one PCR sample each from the patient and a control individual was sequenced. A PCR reaction contains a population of both normal and mutant sequences and therefore analysis of the population would indicate the presence of abnormal sequence. A variant would manifest itself as two tracks with bands at identical positions. A normal control was always sequenced under the same conditions so that the background sequence (from minor populations of PCR products) could be compared and changes eliminated if they occurred in both the control and patient samples.

Chapters Four and Five of this study present two sets of data in which more than one individual possesses the same mutation causing the OI phenotype. The significance of such common mutations is therefore discussed at the end of Chapter Five. Chapters Four through Six discuss glycine by serine substitutions thus the relevance of this class of mutation is discussed fully at the end of Chapter Six. Throughout these chapters, the convention of designating the first glycine of the triple helix as number one has been adhered to.
CHAPTER FOUR

IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN TWO INDIVIDUALS WITH TYPE III OSTEOGENESIS IMPERFECTA

4.1 INTRODUCTION

To date, virtually all mutations which result in osteogenesis imperfecta have been unique to an individual or family. However a few instances have been previously reported in which more than one individual or family share the same mutation and corresponding phenotype; four mutations have been reported as occurring in the COL1A1 gene and two in the COL1A2 gene. Pruchno et al. (1991) identified two probands with type III OI who possess the same new dominant substitution, α1(I)Gly154Arg, and two with type II OI sharing another dominant substitution, α1(I)Gly1003Ser. Wenstrup et al. (1991) identified two families in which affected members with OI type IV all had the same α2(I)Gly646Ser mutation and three individuals with an α2(I)Gly922Ser substitution (Marini et al., 1993a; Szuroloivics et al., 1993; D'Amato et al., 1993) were identified and suffered from type IV OI. All these are instances in which more than one individual or family possessed the same mutation which in each case gave rise to the same severity of OI. However, there are also reports of shared mutations in which the individuals have different severities. Bateman et al. (1992), Mackay et al. (1993a) and Marini et al. (1993a, b) identified three individuals who had an α1(I)Gly352Ser substitution in common. This mutation resulted in lethal through to moderate phenotypes in the affected individuals. In addition, Bateman et al. (1992) and Mottes et al. (1993) reported individuals with moderate/severe and lethal/severe OI phenotypes respectively, who shared an α1(I)Gly415Ser substitution. These are discussed later.

Clinical data regarding probands KO and PJ were made available by Prof. P. H. Byers. KO was born to healthy unrelated parents and has two healthy siblings, both of whom have unaffected children. No other individuals with OI were identified in the family. She had many fractures at birth and was seen to fracture periodically to the date of cell-line production (age 35 years). PJ was identified as having skeletal anomalies in utero at 15 weeks gestation when short limbs were detected. Several subsequent ultrasound examinations throughout the pregnancy confirmed the presence of bowed and possibly fractured long bones. The infant was delivered by Caesarean section for breech presentation (the presence of severe bone deformities means that the foetus cannot move as easily into a 'head-first' position in the uterus). X-rays in the perinatal period demonstrated many features indicative of bone deformity as well as fractures. By age five years he was unable to walk due to multiple and recurrent fractures, was of shortened stature and had a very large head size but was of normal intelligence. These data indicate that although both patients were clinically classified as having OI type III the degree of severity in each case is different; PJ appears to be more severely affected than KO.

Data from the collaborating laboratory indicated that cells from KO and PJ synthesised both normal and abnormal species of type I procollagen and that the abnormal molecules were inefficiently secreted. Cyanogen bromide peptide data indicated that peptides of the abnormal α1(I) chain, amino terminal to residue 822 (that is,
amino terminal to the α1(I)CB6 peptide) all had delayed electrophoretic mobility, suggesting that the mutation was in this region of either the α1(I) or α2(I) chain.

4.2 DETECTION OF A MUTATION IN PATIENTS PJ AND KO

When the α2(I) γ-fragment was analysed by SSCP, bands with altered mobility were detected in the AluVRsal, DdeUSst and MspI digests, corresponding to the 168bp (E/e), 350bp (A/a) and 113bp (E/e) fragments respectively of the named digests (figure 4.1). The smallest region of overlap for the three shifted fragments was identified as a region of 113bp at the 3' end of the fragment (figure 4.2). The fourth digest was DdeVHindII but no band shift was detected in any of the fragments produced. Both patients appeared to be heterozygous for a single mutation since bands of normal mobility were also present in each sample.

In the course of screening, three MspI restriction sites which had been reported by Bernard et al. (1983a) as occurring in the SSCP γ-region of the α2(I) cDNA sequence were not detected in PCR amplified cDNA samples from 18 individuals. The sites reported (codon positions 638, 645 and 747) may, therefore, represent rare variants or DNA sequencing artefacts. The absence of these sites was observed only by restriction analysis — the precise nature of the sequence change was not determined.

4.3 CHARACTERISATION OF THE MUTATION IN PATIENTS PJ AND KO

The cDNA sequence in the region of overlap at the 3' end of the amplified product was determined directly from PCR products from both patients and a control patient who displayed no band shifts for the region. The primer SSCP1A2γR was employed as the sequencing oligonucleotide which yielded the sequence of the antisense DNA strand. In both patient samples, but not that from the control, two nucleotides, G and A, were present simultaneously in the first position of codon 859 of the COL1A2 cDNA (figure 4.3). This resulted in the GGT codon for glycine being altered to an AGT codon for serine. The sequence change destroys a BstUI restriction endonuclease site (recognition sequence 5' CGCG 3'). To confirm the presence of the mutation, a freshly prepared cDNA sample was amplified using SSCP1A2γL and SSCP1A2γR primers and the product cleaved with BstUI. An additional fragment of 125bp was identified only in the patient samples, compatible with the heterozygous mutation (figure 4.4a). The mutation results in the substitution of α2(I) glycine-859 by serine in the products of one allele.

4.4 DETERMINATION OF THE DISTINCTION BETWEEN THE PATIENT SAMPLES

Since only a limited number of OI mutations have been reported to be present in more than one individual, it was necessary to show that the two sample DNAs had not been derived from the same individual which would have lead to an apparently shared mutation. Contamination between two samples may have occurred at the RNA or DNA level in which case the cDNA or genomic DNA samples respectively, would share identity. If,
Figure 4.1  SSCP analysis of patients KO and PJ

Results from the SSCP analysis of the α2 γ-fragment for patients KO and PJ are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples in each digest, lanes 2 and 3 are patients KO and PJ respectively and lanes 1, 4 and 5 are control individuals for the region analysed. Fragments with altered mobility caused by the mutation are indicated by e' in the AluVRsal and MspI digests and by a' in the DdeIStyl digest. Not all fragments are shown. In the DdeIStyl digest, fragments c and d for individual 1, 2, 4 and 5 are faint.
Figure 4.2  Restriction map of the SSCP γ-fragment to illustrate the SSCP overlap pattern for patients KO and PJ

The fragments which yielded a band shift on the SSCP gel are indicated by the bold line on each restriction map (top). The point of mutation as determined by sequence analysis is indicated by an arrow in each case. The peptide fragments corresponding to the PCR fragment are indicated (bottom). R, Rsal; S, Styl; H, HinfI. Relative fragment sizes are indicated, A/a representing the largest fragment. The letters correspond to those on the SSCP figure and those in the text.
Figure 4.3  Sequence characterisation of the mutation in patients KO and PJ

Antisense strand DNA sequencing of double-stranded PCR products showing the region surrounding the mutation in patients KO and PJ and a control individual (N) is presented. The sequence of the sense strand is shown alongside that of the antisense strand. In patients KO and PJ a C to T transition on the antisense strand representing a G to A transition on the sense strand is displayed. This base change converts Gly-859 to serine. The change is not seen in the control individual. The mutant sequence reads 5' GTAACCTCGC 3' as opposed to the normal sequence 5' GTAACCGCG 3'. A BsrUI restriction endonuclease site destroyed by the mutation is indicated by a bar and an asterisk.
on the other hand, the cell lines had been cross-contaminated, both the RNA and genomic DNA samples would share identity. To eliminate the possibility of such cross-contamination, polymorphic sites in both the COL1A1 and COL1A2 gene were examined. In the course of analysing COL1A1 cDNA (Mackay, 1992) it had been determined that KO was heterozygous at an MspI site in the terminal 5' region of the gene (CGG is changed to CGT at codon -120) while PJ was homozygous for the presence of the site. The relevant region of COL1A1 was amplified. Figure 4.4b shows the primers used, the fragment digested and the subsequent products. Oligonucleotides SSCP1A1oL and SSCP1A1oR have the sequences 5' CGGGCTCCTGCTCCTCTTAG 3' and 5' GCCAGGAGCACCAGCAATAC 3', respectively. Detection of the variant by digestion of the product with MspI alone was thought to be difficult. This is because the variant allele generates a fragment 1bp different in length from an MspI fragment derived from a normal allele. The presence of two large similar sized fragments would result in poor separation of variant and normal allele-derived sequences and determination of the variant would have to be made by observing the decrease in intensity of two smaller fragments. However, a double digestion of the DNA with MspI and Styl would restrict the large MspI fragment from a normal individual into two smaller fragments. Following digestion with both enzymes, KO possessed the extra fragment whereas PJ did not thus the digestion indicated that if contamination had occurred, it was not at the RNA or cell line level. To show that the genomic DNA samples had not been cross-contaminated, the state of an Rsal polymorphism in intron 38 of the COL1A2 gene (Grobler-Rabie et al., 1985) was investigated. Amplification of genomic DNA from each patient with the oligonucleotides COL1A1-Rsal primers 1 and 2 (Baker et al., 1991) which amplify a 900bp genomic fragment containing the RFLP, and subsequent Rsal digestion of the product demonstrated that KO was homozygous for the absence of the site whereas PJ was homozygous for its presence (figure 4.4c). These data indicate that each of the patient samples are different and that each possesses a mutation, α2(I)Gly859Ser, which appears to cause the O1 type III phenotype.

4.5 DISCUSSION

The abnormal collagen chains from both patients were found to exhibit overmodification of peptides N-terminal to, and including, α1(I)CB7 suggesting that the mutation lay within the region between codons 551 and 822 of the triple helical region. However the mutation as determined by SSCP and sequence analyses was slightly more carboxyl terminal than predicted by the peptide mapping data. The peptide mapping data were compatible with the mutation being N-terminal to residue 822 of the triple helix inasmuch as the mobility of the α1(D)CB6 peptide from the overmodified α1(I) chains was normal. Residue 822 is the position of the methionyl residue in the α1(I) chain that separates α1(I)CB7 and α1(I)CB6. There is a single lysine residue which lies between the site of mutation and the C-terminal end of the α1(D)CB7 peptide fragment; it occurs at position 855 of the triple helix (figure 4.5) and is the last amino acid of exon 43. There are several possible explanations for the normal mobility of the α1(D)CB6 peptide fragment: (i) the lysine residue is never hydroxylated under any circumstances so that excessive post-translational modification begins in a more amino-terminal location; (ii) the lysine is not usually hydroxylated in the normal α1(I) chain and due to spatial restrictions caused by the presence of the serine residue
Figure 4.4  Restriction enzyme digest to confirm KO and PJ sequence data

(a) Fragments generated from the BamHI digestion of PCR amplified γ-fragment for patients KO, PJ and
controls (C1, C2 and C3) are shown. The presence of the mutation in patients KO and PJ is indicated by an
additional fragment (indicated by an asterisk) originating from the mutant allele.

(b) The map represents a region of COL1A1 amplified by oligonucleotides SSCP1A1xnL and SSCP1A1xnR and
the positions of some of the MspI sites (bars) and Styl sites (bars and 'S'). Not all sites are shown; the jagged
line indicates a region of the amplification product not relevant in this instance. The polymorphic MspI site is
indicated by an asterisk and the Styl site which divides the 225bp MspI fragment (as discussed in the text) is
indicated by an arrow. Fragments generated by an MspI/Styl double digest of KO, PJ and two other individuals
(95 and FW) are shown. The 224bp fragment produced by KO can be seen and is not observed in PJ.

(c) Products generated from the RsaI digestion of an amplified COL1A1 region containing the RsaI RFLP are
shown. Lanes are labelled as follows: M, marker fragments; CHOM+, homozygous control for the presence of
the RFLP; CHET, heterozygous control for the RFLP; CHOM-, homozygous control for the absence of the
RFLP. It can be seen that KO is equivalent to CHOM- whereas PJ is equivalent to CHOM+ at this site. Patients
25 and 95 were included as additional digest samples. The constant fragment of 100bp from this digest is not
shown. All fragments are in base pairs.
\[ \alpha_1(\text{I}) \quad 820 \quad GPWPPGLAGPPGSGREGAPGAEGSPGRDPQA\hat{\text{K}}GDGETGFAAGPP \quad 867 \]

\[ \alpha_2(\text{I}) \quad 820 \quad GAVGSPGSGAPGESGSGREGAPGAEGSPGRDVPQGPH\hat{\text{K}}GERGYPGNIGPV \quad 867 \]

**Figure 4.5** Partial amino acid sequence of the \( \alpha_2(\text{I}) \) SSCP \( \gamma \)-fragment

The amino acid sequence of the regions of the \( \alpha_1(\text{I}) \) and \( \alpha_2(\text{I}) \) polypeptides corresponding to the \( \alpha_2(\text{I}) \) SSCP \( \gamma \)-fragment is illustrated. The M in bold type indicates the N-terminal boundary of the \( \alpha_1(\text{I}) \)C57 peptide fragment. The underlined K indicates the position of the lysine which occurs four residues N-terminal to the substituted glycine (indicated by an asterisk). The single letter amino acid codes follow convention.
in the α2(I) chain (the substitution is only four residues C-terminal to the lysine residue), hydroxylation does not take place in the mutant molecule and (iii) the lysine is hydroxylated in the normal α1(I) chain but due to spatial restrictions as outlined in (ii), glycosylation does not occur. Since serine is a small amino acid, it may be unlikely that it is this alone that causes the spatial restrictions. These restrictions may be brought about by the additive effects of the serine and the Xaa and Yaa position residues in that region. Each of these would indicate that the mutation was N-terminal to residue 822.

The type I collagen peptide sequence has been determined in bovine skin as has the modified nature of different residues (Hofmann et al., 1980). It was determined that the lysyl residue at position 855 in the α1(I) chain is not normally hydroxylated despite lying amino-terminal to a glycine and is thus a candidate for hydroxylation and glycosylation should triple helix formation be delayed. Therefore it is possible that this lysine in the human chain is also not normally hydroxylated. It is uncertain, however, if the mutation four amino acids carboxyl terminal to the residue could disrupt the lysyl hydroxylase binding site or glycosylation site and thus account for failure to modify this residue, or alternatively, if the modification of this residue does not, by itself, lead to a mobility shift in the α1(I)CB6 peptide. Nonetheless, it is clear that peptide mapping has limitations with respect to localisation of mutant residues, both because of ambiguity of the chain involved and because of the relatively large domains identified by protein peptide mapping. These limitations justify the SSCP screening of the whole of the type I collagen genes for mutations and not restricting the search to a specific region.

Information regarding substitutions in the α1(I) chain has indicated that, in general, glycine substitutions towards the C-terminus result in a more severe phenotype than those towards the N-terminus (Starman et al., 1989; Kuivaniemi et al., 1991; Byers, 1993) which is consistent with the mode of helix formation and the α-chain post-translational modification process. Due to the heterotrimeric composition of type I collagen, the phenotype arising from mutations in similar positions on the two α-chains is suggested to be more severe when it occurs in the α1(I) chain. Glycine to serine mutations in the region of the α1(I) chain which corresponds to the region of the α2(I) chain discussed in this chapter, occur at amino acid numbers 844 (OI type III, Puck et al., 1989) and 913 (OI type II, Cohn et al., 1990a) in humans and Stacey et al. (1988) reported a murine α1(I)Gly859Cys synthetic mutation which caused lethal OI. Hence, our data indicate a region on the α2(I) chain where the resulting severity is similar. This is possibly a consequence of the local chain structure, molecular interactions and domain function effects of this region coupled with the effect of the mutation (Chapter Nine; Kuivaniemi et al., 1991; Marini et al., 1993a; Byers et al., 1993).

The mutation found in these two individuals was the only one found in either of the type I collagen genes which could cause the OI phenotype and is, to date, unique to them. It has been discussed that the two patients have differing severities of type III OI, proband PJ appearing to be the more severely affected of the two. It can be speculated that in these cases of differing genotype:phenotype relationships, that the proband with the more severe phenotype may harbour an additional mutation which amplifies the effect of that in the collagen gene or that the proband with the milder phenotype may possess an additional mutation which serves to diminish the effect of the amino acid substitution. A further alternative explanation is that the less severely affected individual may well be a mosaic. If the patient had a de novo mutation during embryogenesis, the defective allele might be present in a
number of different tissue types. In order to investigate the potential mosaic state of this individual, it would be necessary to screen for the mutant allele in various tissues and to observe any of KO’s offspring. The clinical data were compiled on examination of PJ and KO at age 5 years and 35 years respectively. These imply that PJ has a more severe phenotype (or that KO is less severely affected) since an early age patient PJ was already partially immobilised whereas KO was still mobile in her fourth decade. In fact, on a subsequent consultation with the collaborating clinician, it was suggested that the phenotype of patient KO could be better described as a type III/IV.

If such a mutation does exist it may or may not be in one of the type I collagen genes. The entire triple helical coding region of the COL1A1 and COL1A2 genes was analysed using the SSCP technique (Mackay, 1992 and this thesis respectively). If either of these genes in PJ harbours another mutation, it may have escaped detection due to its position and the nature of the resulting SSCP single stranded conformer. These aspects are discussed in Chapter Nine. One prediction that could be made is that if the mutation does reside in a type I collagen gene, it is probably further N-terminal than the Gly859Ser substitution found since the protein data indicate that the mutation is C-terminal to the α(I)CB7 peptide fragment. However, this is only a prediction since from this study it has been seen that protein data are not always accurate. If the mutation does not occur in a type I collagen gene, it is possible that it may be in a gene which in some way affects the collagen biosynthesis pathway e.g. a modifying enzyme, which would not have been detected in a study such as this, or the difference in severity may reflect the effects of other modifying genes, proteins which interact with collagen in the ECM or even in COL1A1. If the latter was true, the mutation has gone undetected by SSCP analysis since no such mutation was characterised for the COL1A1 gene.

The mutation characterised here adds to the data which suggest that glycine by serine mutations do not follow a strict C- to N-terminal trend of decreasing phenotypic severity. In addition, the position of the mutation seen in KO and PJ is interesting in that it occurs just two Gly-Xaa-Yaa repeats N-terminal to the substitution of glycine-865 by serine found by Lamande et al. (1989) in a patient with lethal OI. It would be interesting to observe the severity of OI produced by a serine substitution of the glycine at position 862. This would indicate whether or not this region of the protein represents a critical ‘cut-off’ point between lethal and severe (but survivable) forms of OI, i.e. a domain boundary (see Chapter Six).

Pruchno et al. (1991), have suggested that CpG dinucleotides may be common sites for recurrent mutations in collagen genes, consistent with deamination of a methylcytosine on the antisense strand of DNA. The mutation found in patients KO and PJ is another example of a mutation at such a site. The nature of mutations at sites such as these is discussed in detail in Chapter Nine.

Type III OI was tentatively classified by Sillence et al. (1979) as being a heterogeneous disorder in which both dominant and recessive genotypes could be responsible for the phenotype. This characterisation was refined (Sillence et al., 1984, 1986; Viljoen and Beighton, 1987) and the existence of the type III phenotype was thought to be due to autosomal recessively inherited mutations; a report by Pihlajaniemi et al. (1984), in which an individual had inherited two abnormal proc2(I) collagen alleles with a 4bp deletion, supported this notion. However, several recent reports have suggested that the type III OI phenotype may also be due to autosomal dominant inheritance (Pack et al., 1989; Cohn et al., 1990b; Pruchno et al., 1991; Wenstrup et al., 1991; Molyneux...
et al., 1993). Although instances have been documented in which all types of inheritance can be shown to be responsible for the type III OI phenotype (Byers, 1993), it is widely accepted that most cases are dominantly inherited. The mutation identified in patients KO and PJ is another example of a new dominant mutation resulting in the OI type III phenotype. The major exception to this statement is the number of recessively inherited mutations found in the South African population (Beighton and Versfeld, 1985). This mode of inheritance for severe OI is rare in most populations (Sillence et al., 1986; Thompson et al., 1987) but it may be the most common cause of the phenotype in black South African and Zimbabwe populations (Beighton and Versfeld, 1985; Viljoen and Beighton, 1987; Wallis et al., 1993; section 1.5.1).
CHAPTER FIVE

IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN THREE INDIVIDUALS WITH PERINATAL LETHAL OSTEOGENESIS IMPERFECTA

5.1 INTRODUCTION

It has been seen in Chapter Four that multiple individuals with the same mutation have been reported. In this study, three individuals were found to harbour an identical mutation in the α2 chain of type I collagen resulting in type II OI. All three patients in this study had been screened using SSCP analysis for mutations in the COL1A1 gene which could have accounted for the OI phenotype displayed by each one. No evidence of such a mutation was found and they were subsequently subjected to a similar screening of the COL1A2 gene. Clinical details were available from collaborating laboratories for each patient. Data regarding proband DW were supplied by Dr. Anne De Paepe. She was the only affected child of normal non-consanguineous parents, with three healthy siblings and two healthy half-siblings on the maternal side. The infant died two hours after birth by Caesarean section. She had a number of fractures and displayed many other deformities characteristic of type II OI. Data regarding proband PE were supplied by Prof. Beat Steimann. He was the first child of non-consanguineous parents and there was no family history of OI. The infant died from respiratory failure 44 hours after birth by Caesarean section. Again he displayed clinical and radiological features consistent with lethal OI. Data regarding proband GM02328 had been placed in the NIGMS Human Genetic Mutant Cell Repository by Prof. Hope Punnett who provided clinical information regarding the patient. The infant who had an unaffected sibling was delivered by Caesarean section and died shortly afterwards. He showed multiple fractures and bone anomalies. Type II OI was diagnosed.

Often, cyanogen bromide (CB) peptide mapping data are available for type I collagen from patients with an OI phenotype. However in this case, although available for patients DW and PE, the protein data were not consulted prior to SSCP analysis and mutation characterisation.

5.2 IDENTIFICATION OF A MUTATION IN PATIENTS DW, PE AND GM02328

Upon SSCP analysis, fragments with altered mobility were observed in only one of the four α2(I) SSCP PCR amplified fragments for patients DW and 23 (β-fragment) and two of the four fragments for patient PE (α- and β-fragments). A cDNA template was amplified with oligonucleotides A2A and SSCP1A2PR. SSCP analysis of the β-fragment for patients DW, PE and 23 involved four restriction endonuclease digest combinations. All three patients displayed identical bands with altered mobility in all four digests namely HindII/Styl, HindII/Ddel, HindII/MspI and Alul/Ddel, corresponding to the 194bp (B/b), 239bp (A/a), 321bp (A/a) and 304bp (A/a) fragments respectively of the named digests (figure 5.1). The β-fragment also harbours a common PvuII (or Alul) RFLP reported by Constantinou et al. (1990b). This was detected as an Alul RFLP because of coincidence of an Alul site with the variant PvuII site. However, this known sequence change did not manifest itself as a band with altered
Figure 5.1  SSCP analysis of patients DW, PE and GM02328

Results from the SSCP analysis of the α2 β-fragment for patients DW, PE and GM02328 (23) are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples in each digest, lanes 2, 3 and 4 are patients DW, PE and 23 respectively and lanes 1, 5 and 6 are control individuals for the region analysed. Fragments with altered mobility caused by the mutation are indicated by b' in the HindIII/Styl digest and by a' in the HindII/Ddel, HindII/MspI and AluI/Ddel digests. Arrows indicate spurious PCR products which are identified in order to distinguish them from true band shifts. In the AluI/Ddel digest, the fragments generated as a consequence of the PsrI RFLP are indicated by asterisks in both the native and denatured samples. Not all fragments are shown.
mobility in any of the digest combinations used and thus did not complicate the SSCP analysis of the unknown mutation. The smallest region of overlap was 74bp towards the 3' end of the amplified product (figure 5.2). Each patient appeared to be heterozygous for the mutation since bands of normal mobility were also present in each sample on the SSCP gel.

Primers SSCPIA2oL and A2B were used to amplify a COL1A2 cDNA fragment 933bp in length (α-fragment). SSCP analysis was carried out as above with the appropriate restriction digests. A sequence alteration in patient PE was detected as a band of altered mobility in the PvuII digest and as a loss of a BglII restriction site. The detection of this variant, which is not responsible for the OI phenotype of the proband, is discussed in detail in Chapter Eight but also has relevance in the context of distinguishing the patient materials (see section 5.4).

5.3 CHARACTERISATION OF THE MUTATION IN PATIENTS DW, PE AND 23

The cDNA sequence in the region of overlap in the β-fragment was determined directly from double-stranded PCR amplification products from the patients and a control. The primer SSCPIA2βR was used to sequence the antisense DNA strand. This analysis identified a heterozygous G to A transition in exon 31 in the COL1A2 cDNA from DW, PE and 23 but not in the control sample (figure 5.3). The base change alters the codon GGT to AGT such that glycine-502 is substituted by a serine. The mutation causes the destruction of a PstUI restriction endonuclease recognition site (5' CGCG 3'). To confirm the sequence data, 5 μl of a freshly amplified β-fragment from each of the patients and two control samples were digested with PstUI. An additional fragment of 1026bp was seen in each of the patient samples but not in the controls (figure 5.4) indicating that each patient sample was heterozygous for the PstUI site and therefore possessed the mutation on one allele.

5.4 DETERMINATION OF THE DISTINCTION BETWEEN THE PATIENT SAMPLES

Since cross-contamination of samples at various stages in the analysis could have resulted in identical SSCP band shifts and sequence data in the three individuals, it was necessary to determine whether this had occurred. To eliminate the remote possibility of cross-contamination at the cDNA level, RFLP analyses were carried out. PE possessed a silent mutation resulting in a BglII variant (discussed previously) which was not present in either of the cDNA samples from DW or 23 indicating that PE was a distinct cDNA from the other two patient samples and had not contaminated them (results not shown). PE also had a different haplotype to that of DW and 23 at the PvuII RFLP site in the β-fragment. This can be indirectly determined upon examination of the AluI/Ddel digest pattern of the native fragments in figure 5.1. Since the PvuII recognition sequence coincides with that of AluI, the state of the PvuII RFLP can be inferred from the AluI digest. This indicated that PE was a distinct DNA sample from DW and 23. However, it was not possible to distinguish DW and 23 as being two different cDNA samples since they have the same haplotype for both the BglII variant and the PvuII RFLP. Hence, the identity of the three cell line samples was investigated. If the cell lines from the patients had been cross-contaminated, both
Figure 5.2  Restriction map of the SSCP β-fragment to illustrate the SSCP overlap pattern for patients DW, PE and 23

The fragments which yielded a band shift on the SSCP gel are indicated by the bold line on each restriction map (top). The point of mutation as determined by sequence analysis is indicated by an arrow in each case. The peptide fragments corresponding to the PCR fragment are indicated (bottom). S, StyI; D, Ddel; M, MspI. Relative fragment sizes are indicated, A/a being the largest. The letters correspond to those on the SSCP figures and those in the text.
Figure 5.3  Sequence characterisation of the mutation in patients DW, PE and GM02328

Antisense strand DNA sequencing of double-stranded PCR products showing the region surrounding the mutation in patients DW, PE, GM02328 (23) and a control individual (C) is presented. The sequence of the sense strand is shown alongside that of the antisense strand. In patients DW, PE and 23 a C to T transition on the antisense strand representing a G to A transition on the sense strand is displayed. This base change converts Gly-502 to serine. The change is not seen in the control individual. The mutant sequence reads 5’ GGGGACCGCG 3’ as opposed to the normal sequence 5’ GGGGACTGCG 3’. A BsrUI restriction endonuclease site destroyed by the mutation is indicated by a bar and an asterisk.
Figure 5.4  Restriction enzyme digest to confirm PE, DW and 23 sequence data

Fragments from the BsuUI digestion of patient samples PE, DW and 23 and two control individuals (C1 and C2) are shown. The additional fragment produced due to the mutation in each of the three patients is indicated by an asterisk. All fragment sizes are in base pairs.
the genomic DNA and RNA/cDNA samples would be identical and would, therefore, lead to identical SSCP patterns and hence base sequence. To demonstrate that genomic DNA samples from DW, PE and 23 were distinct, the individuals were digitally mapped at the genomic DNA level using the minisatellite variant repeat PCR (MVR-PCR) mapping technique (Jeffreys et al., 1991). Genomic DNA from each of the three patients was analysed for a-type and t-type repeat units of the human minisatellite MS32 as outlined in section 2.14. This technique relies on the variation at the hypervariable locus D1S8 (probe MS32). This locus displays two types of repeat unit that differ by a single base substitution (G to A) which creates or destroys a HaeIII restriction endonuclease site, as well as length (number of repeats) polymorphism. In principle, no two individuals, save monozygotic twins, share an identical pattern of allelic variation. The MVR-PCR technique exploits the internal variation of this locus. A primer in the 5' flanking DNA of the locus is coupled with an antisense primer specific for the G, or a-type, sequence or for the A, or t-type, sequence. Following the two PCR amplifications, the samples are electrophoresed through an agarose gel, blotted onto a nylon membrane and hybridised with the MS32 probe. The different combinations of a- and t-type repeats can 'identify' an individual. Single alleles are represented by a binary code; an a-type repeat being 1 and a t-type repeat being 2. For genomic DNA where two alleles overlap, a ternary code is generated. Two a-type repeats (aa) are represented by 1, two t-type repeats (tt) by 2 and one of each (al/ta) by 3. There are three other states that can exist. An individual may possess a 'null' or O-type allele i.e. one which contains additional variation 3' to the HaeIII site and has, therefore, failed to amplify. This appears as a missing 'rung' on the MVR coding ladder. The three additional coding states are 4 (aO), 5 (tO) and 6 (OO). The latter would manifest itself as a gap on the ladder.

The five-state code that was generated from genomic DNA from patients DW, PE and 23 demonstrated that each of the samples had a different pattern and could thus be distinguished from each other (figure 5.5). They were, therefore, unique DNA samples.

The mutation, as identified by the RnuUI variant, was detected in each of the individuals from PCR-amplified genomic DNA. Primers MUTA2-31L (sequence 5' GGTCTCCATGGTGAGTTTGG 3') and MUTA2-31R (sequence 5' GACCAGAAGGACCTCGGCTT 3'), in exons 30 and 31 of COL1A2 respectively, were used to amplify a fragment of approximately 1.3kb covering a region from codon 484 to 524 which spans intron 30 of COL1A2 and the mutation in exon 31. The conditions used were 95°C for 1 min, 55°C for 1 min, 70°C for 2 min for 33 cycles. Digestion of crude PCR product from the two patients and three control samples indicated that DW, PE and 23 were heterozygous for the presence of the RnuUI site whereas the controls were homozygous for its presence (Fig. 5.6). These results confirmed the hypothesis that all three patient DNA samples were non-identical and yet harboured the same mutation.

5.5 DISCUSSION

Here data are presented that indicate a heterozygous G to A transition at a CpG dinucleotide has occurred in three different individuals resulting in an α2(I)Gly502Ser substitution in each case. All three
Figure 5.5  MVR-PCR mapping of patients DW, PE and 23

Results from the minisatellite variant repeat PCR (MVR-PCR) typing of total genomic DNA from the three patients are presented. In each case, 'A' represents the a-type repeat unit and 'T' the t-type repeat unit. The five-state code for the superimposed alleles of each patient is given below the map and can be directly compared. Reference marks on the map correspond to those on the codes below and indicate the repeat number with respect to the first on the gel.

The allele codes are denoted by 1 (aa), 2 (tt), 3 (at), 4 (aO) and 5 (tO) where a represents an a-type repeat, t, a t-type repeat and O, a null or O-type repeat.
**DW**

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 40|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 30|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

**PE**

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 40|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 30|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

**23**

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 40|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 30|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 20|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

**DW**

11111333313333113331113341233531133113111121121

**PE**

133313331123111131131111211111113333133

**23**

2211311311233313133113113131333211313313
Figure 5.6  Restriction enzyme digestion to detect the mutation in genomic DNA from the patients PE, DW and 23

Fragments produced from the BstUI digestion of amplified genomic DNA from each of the three patients and control individuals are presented. The region analysed spans the point of mutation and the additional fragment of ~1325bp in patients PE, DW and 23 as compared to the controls (C1 and C2) is indicative of the destruction of the site on one allele and therefore the base defect. Fragments on the left represent marker fragments.
individuals were diagnosed as having type II OI. This is another instance in which multiple individuals have been shown to display the same mutation.

Once the mutation had been identified, protein data from DW and PE were consulted. The data predicted that the mutation would be within the α1(I)CB3 region or the corresponding region of the α2(I) chain i.e. the N-terminal end of α2(I)CB3. The position of mutation as characterised by SSCP and sequence analyses agrees with these data (figure 5.2). Thus it has been shown that in this instance, the RT-PCR-SSCP method coupled with DNA sequencing was sufficient to determine the point and nature of the base change without additional protein data.

Cell line GM02328 has been examined in a previous study carried out by de Wet et al. (1983a). In their report, the cell line was referred to as IMR-2328. By assaying the levels of pro-α1(I) mRNA relative to the pro-α2(I) mRNA and the ratio of rates of synthesis of pro-α1(I) and pro-α2(I) chains by cultured fibroblasts, they concluded that the variant causing the lethal OI phenotype must involve a mutation which affects the expression of the pro-α2(I) gene. No mutation data were presented. The study carried out here would indicate a mutation in the α2(I) gene which results in an abnormal α-chain being synthesised which is incorporated in the procollagen molecule. Obviously these two sets of data are in conflict unless the proband GM02328 harbours two mutations. One of these mutations could affect the expression of the gene and if expression is not abolished completely, as is implied by de Wet et al. (1983a), the synthesised chains which then may contain the Gly502Ser substitution can be incorporated into the mature α-chain.

5.5.1 Multiple occurrence of mutations

There have been several reports which have indicated that a mutation is shared by more than one individual or family. There have been four such instances in the α1(I) chain and three in the α2(I) chain including that documented in Chapter Four. It was suggested by Pruchno et al. (1991) that CpG dinucleotides within the coding region of either type I collagen gene could be indicative of 'hot-spots' for mutations (provided that they occur at a non-intron-exon boundary Yaa-Gly amino acid pair). Of the mutations characterised to date in more than one individual or family, including those described in this thesis, all occur at such dinucleotide sequences. The mechanism of mutation at these sites is discussed in full in Chapter Nine.
CHAPTER SIX
IDENTIFICATION AND CHARACTERISATION OF A GLYCINE TO SERINE MUTATION IN A PATIENT WITH TYPE IV OSTEOGENESIS IMPERFECTA

6.1 INTRODUCTION
The patient in this study (SC) was diagnosed as having type IV or moderate 01. CB peptide data supplied by the collaborating laboratory indicated that overmodification of the procollagen began at CB8 on the α1 chain or an equivalent region on the α2 chain of type I procollagen. Once again, screening of the COL1A1 gene for sequence defects had yielded no such alterations (Mackay, 1992).

6.2 IDENTIFICATION OF THE MUTATION IN PATIENT SC
Following SSCP analysis of the α2(I) α-fragment, bands with altered mobility were detected in the 360bp (A/a), 159bp (C/c) and 246bp (B/b) fragments of the PvuII, BglII and AluI digests respectively (figure 6.1). A band with altered mobility was not detected in fragments from the NcoI/PvuII digest. The smallest region of overlap was 107bp towards the 3' end of the amplified fragment (figure 6.2). Analysis of the remaining three SSCP PCR fragments yielded no further band shifts.

6.3 CHARACTERISATION OF THE MUTATION IN PATIENT SC
A region of cDNA surrounding that expected to harbour the mutation in patient SC was cloned. The amplified α-fragment was digested with PstI which cuts within the collagenous sequence and EcoRI which cuts within the 5' extension sequence of primer A2B. The resulting 703bp fragment (figure 6.3a) was ligated to identical sites within the polylinker of plasmid vector pTZ19R. Following transformation (section 2.7.3.1) potential recombinant colonies were used to generate bacterial cultures. Sequence analysis of DNA isolated from the bacteria was performed using an internal oligonucleotide specific for the coding region of α2(I) collagen sequence (A2C; sequence 5' GGTAATCCTGGAGCAAACGG 3'; Molyneux, 1991). The position of this primer is illustrated in figure 6.3a. Analysis revealed a G to A transition in exon 19 in three of the eight samples sequenced. This base change causes a GGT glycine codon at position 238 to be altered to AGT encoding serine in the heterozygous individual (figure 6.3b). The mutation destroys a BstUl restriction site. Digestion of freshly prepared cDNA samples from the patient and control individuals yielded the bands expected from the digestion of such individuals (results not shown).
Figure 6.1  SSCP analysis of patient SC

Results from the SSCP analysis of the α2- Fragment for patient SC are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples in each digest, lane 3 is patient SC and lanes 1, 2, 4 and 5 are control individuals for the region analyzed. Fragments with altered mobility caused by the mutation are indicated by a' in the PvuII digest, by c' in the BglII digest and by b' in the AflII digest. Not all fragments are shown.
Figure 6.2  Restriction map of the SSCP α-fragment to illustrate the SSCP pattern for patient SC

The fragments which yielded a band shift on the SSCP gel are indicated by the heavy line on the restriction maps (top). The point of mutation as determined by sequence analysis is indicated by an arrow in each case. The peptide fragments corresponding to the PCR fragment are indicated (bottom). P, Pvull. Relative fragment sizes are indicated, A/a being the largest. The letters correspond to those on the SSCP figures and those in the text.
Figure 6.3  Characterisation of the mutation in patient SC

(a) An AluI restriction map of the SSCP α-fragment is shown. The hatched box represents the 13bp tag sequence of the oligonucleotide A2B (not to scale). The 703bp fragment cloned for sequencing is indicated as is the position of the sequencing primer, A2C, with respect to the cloned fragment.

Restriction sites are labelled as follows: P, PstI; E, EcoRI.

(b) DNA sequencing of the sense strand of cloned PCR products showing the region surrounding the mutation in patient SC is presented. Variant sequence (V) differs from normal sequence (N) in that a G to A transition has taken place. This base change converts glycine-238 to serine. The mutant sequence reads

5' CCGCAGTATCCT 3' as opposed to the normal sequence 5' CCGCGGTATCCT 3'. A BsuRI restriction endonuclease site destroyed by the mutation is indicated by a bar and an asterisk.
(a)

703 bp

(b)

5' GC GC GC C

3' T C C T G T A G C

5' A C G C C C

3' G A T C
6.4 DISCUSSION

The protein data that were provided for this patient indicated that the mutation should be within the \( \alpha 1(II)CB8 \) domain or the corresponding region on the \( \alpha 2 \) chain. This region spans residues 123 to 401. The substitution occurring at position 238 in proband SC is in agreement with this prediction.

This and the preceding two chapters have documented three different glycine by serine substitutions in the \( \alpha 2 \) chain of type I collagen. These result in the range of OI phenotypes expected from an included mutant collagen chain and add considerably to the number of substitutions characterised in the \( \alpha 2 \) chain. In addition, the Gly-502 and Gly-859 substitutions have been shown to be potential highly mutable regions due to more than one individual possessing such an altered residue. Also, the Gly-238 substitution is the most N-terminal substitution of any amino acid found in the \( \alpha 2 \) chain. The moderate phenotype associated with this change suggested in advance that the mutation might be situated towards the 5' end of the gene.

As has been discussed in Chapter One, domains and co-operative blocks may exist along type I collagen. From the accumulation of data concerning substitution of glycine by serine and cysteine it can be seen that these two amino acids do not adhere to a simple gradient model since lethal substitutions occur N-terminal to non-lethal ones. Since these substitutions appear to be the most variable in terms of the resulting OI phenotype with respect to position along the \( \alpha \)-chain, they can be used to predict domain boundaries along each of the collagen chains. It is most probable that the X and Y position amino acids determine the limits of the boundaries since these are the variable factors between different \( \text{Gly-Xaa-Yaa} \) repeat motifs. Since the composition of the \( \alpha 1 \) and \( \alpha 2 \) chains differs in these residues, it might be expected that the domains along the \( \alpha 1 \) chain may differ in position from those on the \( \alpha 2 \) chain. Figure 6.4 illustrates the domains predicted from known substitutions including the three glycine by serine substitutions characterised in this thesis. The domain boundaries were assigned from the C-terminal end of each chain taking into consideration substitutions by both amino acids. If a single residue position did not agree with the region depicted, it was assumed to be an anomaly (see below). However, the presence of two adjacent substitutions was treated as a domain in its own right regardless of whether these substitutions were by the same amino acid or not. Obviously the boundaries can only be assumed to be tentative since relatively few substitutions have been characterised with respect to the number of potential substitutions in each chain and a full investigation of the nature of the amino acids in the X and Y positions was not carried out. The complete picture could only be drawn if all glycine substitutions could be identified. In this situation the theory of domains may well be disproved due to a particularly complicated interspersion pattern of lethal and non-lethal OI-causing substitutions.

Within the currently predicted domains (figure 6.4), some residues do not conform to the phenotype predicted. On the \( \alpha 1(II) \) chain, the mutation at position 415 results in a lethal phenotype and yet is nested in the middle of a 'non-lethal' domain. The reverse is true for glycine substitutions at residues 589 and 973. Residue 1003 causes non-lethal OI and is at the end of a 'lethal' domain. All of these residues have been substituted by serine. On the \( \alpha 2 \) chain, the substitution causing non-lethal OI at position 859 is towards the end of a 'lethal' domain. This is also a serine substitution.

Such anomalies could be indicative of two factors. One explanation could be the existence of 'sub-domains' whereby 'non-lethal' domains are divided into regions of mild, moderate and severe phenotypes.
Figure 6.4 Predicted domains along the type I collagen chains

All known cysteine and serine substitutions for glycine on the α1 and α2 chains of type I collagen are indicated. The positions of mutations that cause lethal (L) and non-lethal (N) OI phenotypes are shown below each map. From this information possible domain boundaries have been postulated for each type I collagen chain and are indicated by arrows on the bold lines (DOM). Since the α-chains associate first at the C-termini the domains were assigned from this end towards the N-terminus. The phenotypes generated from substitutions by both cysteine and serine were considered simultaneously when proposing a domain type. The domains are termed non-lethal (N) if the majority of mutations within the region cause a non-lethal OI phenotype and lethal (L) if the majority of mutations within the region cause the lethal OI phenotype. A possible different pattern in the α2 chain is indicated by arrows and lower case letters. Bars marked with an asterisk indicate mutations characterised in this thesis. Bars marked with a cross indicate positions which do not conform exactly to the domains predicted. These are the only positions numbered and are discussed in the text.
Figure 6.5a illustrates possible sub-domains of the N-terminal ‘non-lethal’ domain of the α1(I) chain. This may occur due to other influencing factors such as local amino acid effects. Where lethal phenotypes occur in a ‘non-lethal’ domain, the same principle may be true. The extreme C-terminal ‘lethal’ domain on the α1(I) chain has an alternating pattern of severe and lethal phenotypes which may indicate sub-domains. A ‘non-lethal’ domain of the α1(I) chain has been expanded to illustrate this point (figure 6.5b). However, in this instance, the aberrant amino acid can be substituted by a cysteine and only lead to a moderate/severe phenotype suggesting that the effect of a cysteine is less devastating than a serine. The second explanation that can be invoked is the possibility of the patient harbouring another mutation. The patient possessing the α1(I)Gly415Ser mutation may have an additional defect which amplifies the phenotypic outcome.

The substitution of glycine-859 in the α2 chain by serine is the one documented in Chapter Four. Two individuals possess this mutation. In one it causes a type III phenotype and in the other a type III/IV phenotype. Due to the two neighbouring known substitutions it is difficult to say whether this or the residue at position 706 is in a wrongly classified domain. Figure 6.4 illustrates an alternative domain pattern with the existence of a narrow lethal ‘window’. As suggested in Chapter Four, the phenotype caused by a glycine by serine substitution at position 862 would be interesting information. In this circumstance it would be useful in determining a more precise domain boundary.
Figure 6.5  Predicted sub-domains for selected domains of the
α(I) chain of type I collagen

Selected domains from figure 6.4 are illustrated in expanded form and not to scale with respect to each other. (a) The N-terminal 'non-lethal' domain of the α(I) chain and (b) the central 'non-lethal' domain from the α(I) chain. In each case the α(I) domain map is indicated above the sub-domain maps; dashed lines indicate the region amplified. Arrows indicate the main domain boundaries and triangles indicate the possible sub-domain boundaries for each region. Substitutions are indicated by bars. C, cysteine substitutions; S, serine substitutions; D, domains. Sub-domains are termed in the same manner as individual substitutions i.e. S, severe; Mo, moderate; Mi, mild. Domains are termed lethal (L) or non-lethal (N).
CHAPTER SEVEN

THE IDENTIFICATION AND CHARACTERISATION OF A NOVEL GLYCINE TO GLUTAMIC ACID MUTATION IN A PATIENT WITH PERINATAL LETHAL OSTEONEGENESIS IMPERFECTA

7.1 INTRODUCTION
To date, the substitution of glycine by glutamic acid has not been observed in reported cases of OI. In this study such a mutation was characterised in a patient with lethal OI. Clinical and biochemical data were available from the collaborating laboratory. Proband R was the first child born to phenotypically normal non-consanguineous parents and displayed characteristic features of the perinatal lethal form of OI (soft calvarium, small thorax, short limbs). The infant died shortly after delivery. Cells from the affected infant synthesised and secreted a diminished amount of normal type I procollagen molecules and retained within the cells a species of type I procollagen in which both the procI(I) and proc2(I) chains had delayed electrophoretic mobilities. Very few of the abnormal molecules appeared to be secreted. The cyanogen bromide peptides obtained from the amino-terminal end of the triple helix of the abnormal α1(I) chains within the cells also had delayed electrophoretic mobilities, compatible with overmodification in the domain amino-terminal to residue 401 of the triple helix, suggesting that the mutation lay between residues 123 and 401 of the triple helix (exons 13-25) of either type I collagen gene. Parental fibroblasts synthesised only normal type I procollagen molecules.

7.2 IDENTIFICATION OF THE MUTATION IN PATIENT R
No abnormalities were detected in the entire COL1A1 cDNA sequence using SSCP analysis (Mackay, 1992). In contrast, the COL1A2 fragment amplified by primers A2A and SSCP1A2B contained electrophoretic variants. Following first strand cDNA synthesis, PCR amplification was carried out and SSCP analysis performed. Fragments with altered mobilities were observed in the HindIIIdel (C/c), HindISyl (C/c) and HindIIMspI (B/b) digests (figure 7.1) corresponding to the 189bp non-denatured HindII fragment in each instance (figure 7.2). A fragment with altered mobility was not observed in the fourth digest (AluVDdel). In each instance both a normal and abnormal fragment was seen, consistent with heterozygosity for a mutation.

7.3 CHARACTERISATION OF THE MUTATION IN PATIENT R
The sequence of the sense strand of DNA from patient R and a control individual was determined directly from a cDNA PCR product. A collagen sequence specific internal oligonucleotide (SEQA2-21; sequence 5’ CTCCAGGACCTCCTGGGCTG 3’) was used. Analysis revealed a G to A transition in the GGA codon encoding a2(I)Gly-343 in one allele which creates the GAA codon encoding a glutamic acid (figure 7.3a). The mutation destroys an AvaII restriction site; digestion of cDNA samples from the patients’ cells produced the
Results from the SSCP analysis of the α2 β-fragment for patient R are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples in each digest, lane 1 is patient R and lanes 2, 3 and 4 are control individuals for the region analysed. Fragments with altered mobility caused by the mutation are indicated by c' in the Hinfl/Dele and Hinfl/SstI digests and by b' in the Hinfl/MspI digest. Not all fragments are shown.
Figure 7.2  Restriction map of the SSCP β-fragment to illustrate the SSCP pattern for patient R

The fragments which yielded a band shift on the SSCP gel are indicated by the bold line on each restriction map (top). The point of mutation as determined by sequence analysis is indicated by an arrow. The peptide fragments corresponding to the PCR fragment are indicated (bottom). S, Syl; D, Ddel; M, Mspl. Relative fragment sizes are indicated, A/a being the largest. The letters correspond to those on the SSCP figures and those in the text.
Figure 7.3  Characterisation of the mutation in patient R

(a) Sense strand DNA sequencing of double-stranded PCR products showing the region surrounding the mutation in patient R and a control individual (N) is presented. The analysis reveals a G to A transition in the patient not seen in the control individual. This base change converts Gly-343 to glutamic acid. The mutant sequence reads 5' CGAGAACCT 3' as opposed to the normal sequence 5' CGAGGACCT 3'. An AvaII restriction endonuclease site destroyed by the mutation is indicated by a bar and an asterisk.

(b) The sequence data were confirmed by digestion of amplified cDNA product from R and three control individuals (C1, C2 and C3) with AvaII. The additional band of 162bp (as indicated by an asterisk) accompanied by a loss of intensity of fragments 117bp and 54bp in length can be seen for R but not for the control individuals. This is indicative of the mutation occurring in one allele. Not all fragments are shown.
fragments expected from such a variant that was not present in several controls (figure 7.3b). The position of the mutation is consistent with the region predicted by the CB peptide mapping data. SSCP analysis of the remaining \( \alpha_2(I) \) cDNA sequence from the triple helical region revealed no further fragments with altered mobilities. Skin fibroblast cell lines from the parents of proband R were available. In order to determine how the mutation had occurred (parental germline/somatic mosaicism or de novo mutation) genomic DNA was isolated from parental and proband cultured fibroblast cells. A region of genomic DNA, approximately 1.2kb in length, was amplified with primers GLU-L and GLU-R which have the sequences 5’ GGCCCTCCTGTGGTAGTCGTGG 3’ and 5’ TCCAGCGGGGCCGATATTTC 3’ respectively and reside in exons either side of intron 23. The conditions of amplification were 95°C for 1 min, 55°C for 1 min, 70°C for 2 min for 33 cycles. Digestion of crude amplified product from both parents, the proband and control individuals demonstrated that only the affected individual possessed the mutation as indicated by the heterozygosity of the AvelI site. The parents and control individuals were homozygotes for the presence of the restriction site (figure 7.4). The parents of R do not appear to be mosaic in their fibroblasts for the mutation harboured by the proband. This does not preclude them from being mosaics it merely suggests that they are not.

7.4 DISCUSSION

Point mutations in the first position of the GGN glycine codon would result in codons for arginine, serine, cysteine or a stop codon, whereas those in the second position would result in codons for alanine, valine, aspartic acid or glutamic acid. The previous failure to identify substitution of a glycine residue by glutamic acid in the chains of type I collagen possibly reflects two factors: (i) the paucity of the GGG codon for glycine — 3.3% in the \( \alpha_1 \) mRNA and 4.5% in the \( \alpha_2 \) mRNA of the triple helical domain — or (ii) the relative infrequency of second position mutations in glycine codons which might counteract the relatively high frequency of the GGA codon for glycine with respect to the GGG codon — 18.1% in the \( \alpha_1 \) mRNA and 22.6% in the \( \alpha_2 \) mRNA of the triple helical domain. This means that since the occurrence of such mutations is rarer than others, the phenotypic consequences of substitutions of glycine by glutamic acid are difficult to predict. However, several cases of substitution by aspartic acid have been documented in both the \( \alpha_1(I) \) and \( \alpha_2(I) \) collagen chains. In each instance, the \( \alpha_1 \) phenotype resulting from the mutation was the lethal variant. In the \( \alpha_1(I) \) chain, glycine substitutions by aspartic acid occur at amino acid positions 97 (Lightfoot et al., 1992), 541 (Zhuang et al., 1991), 559 (Byers, 1990), 673 (Kuivaniemi et al., 1991) and 883 (Cohn et al., 1990b). In the \( \alpha_2(I) \) chain similar substitutions occur at positions 547 (Bonadio et al., 1988), 580 (Niyibizi et al., 1992), 700 (Gomez-Lina et al., 1993), 805 (Grange et al., 1990), 907 (Baldwin et al., 1989) and 976 (Byers, 1990). It is possible that both glutamic acid and aspartic acid, when incorporated into the mature \( \alpha \)-chain, disrupt the inter- and intra-molecular interaction of the mature collagen molecule in a similar manner due to the acidic nature of the side chains of both. It is possible that the higher occurrence of Gly by Asp compared to Gly by Glu substitutions is due to the higher occurrence of the GGT codon than either the GGA or GGG codons. This in addition to the similar frequencies of the GGA and GGC codons may prejudice the outcome of a G to A transition in the second position of a glycine codon.
Figure 7.4  The occurrence of the mutation in patient R

(a) Genomic DNA was amplified with primers GLU-L and GLU-R. The approximately 1170bp product was digested with AvaII. Products of the digestion are shown for the father (F), proband (P), mother (M) and a control (C). The 91bp fragment indicative of mutation is present in the proband but neither of the parents nor the control individual. The 46/45bp fragment in sample F is faint. The three bands whose sizes can only be approximated are derived from the intronic sequence.

(b) An AvaII restriction map of the amplified product. Primers are represented by solid arrows. A, AvaII restriction site. The AvaII site destroyed by the mutation is indicated by an asterisk. Diagonal lines represent intron sequence of approximately 1040bp. Dashed lines represent the exons (E 23 and E 24). Known fragment sizes are indicated. The positions of the AvaII sites within the intron were not determined.
The position of the glutamic acid substitution reported here (Gly-343) is more N-terminal on the α2(I) chain than any of the documented glycine by aspartic acid substitutions. However, no assumptions can be made on the phenotype that might arise from similar glycine by glutamic acid substitutions further C-terminal, although from the observed phenotype pattern resulting from substitution by aspartic acid (which has a shorter side chain and would therefore be expected to be less disruptive), it can be postulated that glutamic acid substitutions might result in a lethal phenotype wherever they occur along the triple helical domain.

In virtually all instances, OI type II results from dominant de novo mutations in which a glycine codon in the domain encoding the triple helix in one of the two type I collagen genes (COL1A1 or COL1A2) is altered to encode a bulkier amino acid; the new amino acid interfering with the winding of the three proc-helices into the procollagen molecule. Here the detection of a heterozygous G to A transition in COL1A2 is described. This mutation causes the glycine at position 343 of the triple helix of α2(I) to be substituted by glutamic acid and resulted in the lethal variant of OI in the proband.
CHAPTER EIGHT

VARIANTS IN THE COL1A2 GENE DETECTED DURING SSCP MUTATION DETECTION ANALYSIS

8.1 INTRODUCTION
It was not unexpected that the SSCP mutation detection system should detect non-OI-causing sequence changes. Five had been found in the COL1A1 gene during screening (Maclntyre, 1992). In this study two sequence changes were found in the COL1A2 gene which could not be the cause of the OI phenotype displayed by the proband. The single base changes that were detected occurred in the third position of the codon and thus did not cause an amino acid change; these were therefore classified as silent mutations.

8.2 A BglII VARIANT IN PATIENT PE
8.2.1 Detection of the mutation
The proband PE who was diagnosed as having the perinatal lethal form of OI, has been discussed previously in Chapter Five since a mutation was found which was the cause of the OI phenotype. Further band shifts were detected upon SSCP analysis of the α-fragment; shifts were seen in the largest fragment of the PvuII digest (360bp; A/a) and the 246bp band of the AluI digest (B/b) (figure 8.1). The region of overlap was 246bp (figure 8.2a). It was not known whether the individual was a compound heterozygote i.e. possessed two mutations, the combined effect of which was the type II OI phenotype displayed. A similar phenomenon had been observed (de Wet et al., 1983b, 1986; Tromp and Prockop, 1988) in which an individual had two altered alleles for the proα2(I) chains of type 1 procollagen. The mutations combined to produce a lethal phenotype. One of the mutations produced a non-functional allele i.e. mRNA synthesis occurred but proα2(I) chain synthesis did not. The mutation in the other allele was an A to G transition in a universal consensus sequence RNA splice site. This caused inefficient splicing of the RNA from the last codon of exon 27 to the first codon of exon 29. For patient PE, as well as the shift, an additional band was observed in the native BglII digest sample which was accompanied by the decreased intensity of the 72bp and 172bp fragments (figure 8.2b). This was indicative of a polymorphic state of the most 3' BglII site within the amplified fragment.

8.2.2 Characterisation of the mutation
In order to sequence the mutation a stretch of amplified DNA encompassing the region of interest was cloned. The same fragment which was cloned for patient SC (Chapter Six) was used for PE. Figure 8.3a shows the SSCP α-fragment. The position of the sequencing primer is indicated with respect to the polymorphic BglII site as well as the fragment cloned for sequencing purposes. Using a collagen specific coding primer (A2C; Chapter Six) situated within the cloned fragment, the mutation was shown to be a C to T transition within codon 226 in exon 19.

73
Figure 8.1 SSCP analysis of patient PE

Results from the SSCP analysis of the α2-fragment for patient PE are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples in each digest, lane 1 is patient PE and lanes 2, 3 and 4 are control individuals for the region analysed. The BglII digest shows the variant BglII site as indicated by an arrow and a V in the native sample and arrows and v in the denatured sample. Fragments with altered mobility caused by the mutation are indicated by α' in the PvuII digest and by b' in the Alul digest. Not all fragments are shown.
Figure 8.2  Localisation of the variant $Bgl$ site in the $\alpha$-fragment of patient PE

(a) The fragments which yielded a band shift on the SSCP gel are indicated by the bold line on each restriction map. Relative fragment sizes are indicated, A/a representing the largest fragment. The letters correspond to those on the SSCP gel. The variant $Bgl$ site is indicated by an asterisk.

Restriction sites are identified by $P$, $PvuII$.

(b) The $Bgl$ variant site of PE is illustrated. C1, C2 and C3 are control individuals. The additional fragment of 244bp in the patient (as indicated by an asterisk) is accompanied by a loss of intensity of fragments 172bp and 72bp in length. All fragment sizes are in base pairs.
Figure 8.3  Characterisation of the mutation in patient PE

(a) A BglII restriction map of the SSCP α-fragment is shown. The hatched box represents the 13bp tag sequence of the oligonucleotide A2B (not to scale). The 703bp fragment cloned for sequencing is illustrated as is the position of the sequencing primer, A2C, with respect to the cloned fragment. The variant BglII site is indicated by an asterisk.

Restriction sites are labelled as follows: P, PstI; B, BglII; E, EcoRI.

(b) DNA sequencing of the sense strand of cloned PCR products showing the region surrounding the mutation in patient PE is presented. Variant sequence (V) differs from normal sequence (N) in that a C to T transition has taken place (as indicated by an asterisk). This base change converts codon 226 from GGC to GGT. The mutant sequence reads 5' CCCGGTGTTGCT 3' as opposed to the normal sequence 5' CCCGGCGTTGCT 3'.
The sequence change was observed in three out of eight clones. This caused the codon GGC to be altered to GGT and since the amino acid is not altered, this mutation is described as silent.

8.2.3 Frequency of the \textit{BglII} variant

Of the 17 patients screened for mutations in the \( \alpha \)-fragment, only PE demonstrated the presence of the \textit{BglII} variant. Since the variant has such a low frequency, it is unlikely that it will be of any use in RFLP linkage studies.

8.3 AN \textit{AluI} RFLP IN THE POPULATION OF PATIENTS

8.3.1 Detection of the mutation

A \textit{PvuII} RFLP had been characterised by Constantinou \textit{et al.} (1990b). The base change was an A to C transversion at residue 392 of the \( \alpha \)-chain. This did not alter the proline at this position. In the course of analysing the \( \beta \)-fragment by SSCP, this variant was detected as an \textit{AluI} RFLP since the recognition sequences of these two enzymes coincide. This RFLP did not manifest itself as a band shift in any of the digests but was detected in the native samples as an additional band with a corresponding loss of intensity of the 207bp and 142bp fragments in the \textit{AluI/DdeI} digest (results not shown).

8.3.2 Frequency of the RFLP

In this study it was seen that for a total of 15 patients, two were homozygous for the absence of the \textit{AluI} site, five were homozygous for its presence and eight were heterozygous for the site. The heterozygosity of the RFLP in this study is 53.3\% which is almost in agreement with the 45\% quoted by Constantinou \textit{et al.} (1990b). However, no comparisons can be made between the two populations since Constantinou \textit{et al.} (1990b) do not quote actual numbers of individuals for the different states of the variant site.

8.4 A SILENT MUTATION IN PATIENT 45

8.4.1 Detection of the mutation

Patient 45 was classified as having a non-lethal form of OI. Bands with altered mobility were observed in the analysis of the SSCP 8-fragment. A shift was seen in the 395bp fragment of the \textit{HinclI/RsaI} digest only (figure 8.4a).

8.4.3 Characterisation of the mutation

In order to determine the nature of the mutation, the 116bp and 332bp \textit{ApaI} fragments from the amplified region (figure 8.4b) were cloned into phosphatase-digested, \textit{ApaI}-digested \textit{pBluescript KS}\(^\text{\textregistered}\). Of the eleven 116bp \textit{ApaI} clones which were sequenced using the M13 reverse primer, seven displayed a C to T transition in the third position of codon 955 in exon 47 (figure 8.4c). The mutation caused the codon GGC to be altered to GGT and was
Figure 8.4  Characterisation of the mutation in patient GM02645

(a) Results from the SSCP analysis of the α2-8-fragment of patient GM02645 (45) and control individuals are presented. Native (n) and denatured (d) fragments are indicated by upper and lower case letters respectively. For native and denatured samples, lane 1 is patient 45 and lanes 2, 3, 4 and 5 are control individuals for the region analysed. The fragment with altered mobility caused by the mutation is indicated by a'. Not all fragments are shown. Spurious PCR fragments are indicated by arrows to avoid confusion in determining band shifts.

(b) The digestion fragment which displayed a band with altered mobility following SSCP analysis is shown on the bottom line with respect to fragments generated from an Apal/Ddel digest of the same region. The 332bp and 116bp Apal fragments were cloned for sequencing purposes.

Restriction sites are labelled as follows: A, Apal; D, Ddel; H, HincII; R, Rsal.

(c) Sequence results of the sense strand of DNA from recombinant plasmids harbouring the 116bp Apal fragment from (b) are presented. The variant sequence (V) differs from the normal sequence (N) in that a C to T transition has occurred altering codon 955 from GGC to GGT. The mutant sequence reads 5' GGTTCCGTG 3' as opposed to the normal sequence 5' GGCTCCGTG 3'.
thus classed as a silent mutation. This variant was only detected in this patient and is, therefore, probably a rare variant: this combined with the fact that the base change does not alter a known restriction enzyme restriction site means that the variant will not be of use in linkage studies. This mutation did not cause the OI phenotype of the proband.

3.5 DISCUSSION

Of the 17 patients screened for a mutation in the COL1A2 gene which could be the cause of the OI phenotype, two possessed silent mutations (PE and 45). The mutation causing the phenotype of PE has been characterised (Chapter Five) whereas that of 45 has not. Band shifts in the other three SSCP fragments were not observed for patient 45 and therefore the mutation underlying the OI phenotype was not identified.
CHAPTER NINE

COMMENTS ON THE SSCP ANALYSIS RESULTS AND FURTHER MUTATION ANALYSIS

9.1 INTRODUCTION

This chapter attempts to draw together information amassed during the examination of the patient samples by SSCP analysis of the type I collagen genes as well as an investigation into suitable alternative radionucleotides to be used in SSCP analysis. A further screening strategy for OI mutations which was investigated is also discussed.

9.2 POSITIONS OF THE CHARACTERISED SUBSTITUTIONS WITH RESPECT TO REPORTED DATA

In this study mutations that resulted in the OI phenotype were identified in seven individuals although only four different mutations were responsible for the phenotypes. Two of these mutations caused the lethal form of OI, one caused the severe form and the remaining one resulted in moderate OI. Two of the mutations occurred in multiple individuals and the position of the base change at a CpG dinucleotide in each case adds to the data suggesting that such sequences are more easily mutable than others. The first glycine by glutamic acid substitution in type I collagen was identified and the resulting OI phenotypic similarity between substitution of glutamic and aspartic acids was speculated upon. In addition, two base changes that resulted in a silent sequence variant were characterised and a previously characterised RFLP was detected during SSCP analysis. The novel data are summarised in table 9.1. Of the patients in which an OI mutation was found, protein data were available for six and in four cases the position of the mutation was consistent with that predicted from examination of the over-modification of the α-chains. However, in the remaining two there was a discrepancy between the position of the mutation as determined by SSCP and sequence analyses and that predicted by protein analysis. This is an indication that cyanogen bromide peptide mapping has its limitations although in many cases it is a useful method for the general localisation of a defect. In the case of the mutation shared by three individuals, protein data were not consulted prior to mutation identification. The identification and characterisation of this mutation demonstrates that SSCP analysis coupled with sequencing analysis may be sufficient to detect a base change in some instances. The positions of the substitutions characterised in this thesis, with respect to known data, in addition to the variants found are indicated in figure 9.1.

9.3 AN ESTIMATION OF THE EFFICIENCY OF THE SSCP ANALYSIS TECHNIQUE

Following the SSCP analysis of both the COL1A1 and COL1A2 genes, it is possible to estimate the efficiency of the system for the detection of OI mutations. In total, the type I collagen genes of 28 patients were investigated in this manner. This number included two patients with previously characterised mutations and one
### Table 9.1 Mutations characterised by SSCP and sequence analysis.

The mutations that were found by SSCP and sequence analysis are presented. Where the table is blank, no mutation was identified. The third column states whether the mutation resulted in an OI phenotype or not.

<table>
<thead>
<tr>
<th>PATIENT IDENTIFICATION</th>
<th>MUTATION CHARACTERISED</th>
<th>RESULTS IN OI PHENOTYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM02328</td>
<td>α2(I)Gly502Ser</td>
<td>YES</td>
</tr>
<tr>
<td>NA</td>
<td>α2(I)Gly502Ser</td>
<td>YES</td>
</tr>
<tr>
<td>PE</td>
<td>α2(I)Gly502Ser</td>
<td>YES</td>
</tr>
<tr>
<td>FW</td>
<td>α2(I)Gly226Gly</td>
<td>NO</td>
</tr>
<tr>
<td>R</td>
<td>α2(I)Gly343Glu</td>
<td>YES</td>
</tr>
<tr>
<td>DW</td>
<td>α2(I)Gly502Ser</td>
<td>YES</td>
</tr>
<tr>
<td>BF</td>
<td>α2(I)Gly859Ser</td>
<td>YES</td>
</tr>
<tr>
<td>PJ</td>
<td>α2(I)Gly859Ser</td>
<td>YES</td>
</tr>
<tr>
<td>PIR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>α2(I)Gly859Ser</td>
<td>YES</td>
</tr>
<tr>
<td>SC</td>
<td>α2(I)Gly238Ser</td>
<td>YES</td>
</tr>
<tr>
<td>CF-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM02573</td>
<td>α2(I)Gly955Gly</td>
<td>NO</td>
</tr>
<tr>
<td>GM01093</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM02695</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GM02645</td>
<td>α2(I)Gly955Gly</td>
<td>NO</td>
</tr>
</tbody>
</table>
Figure 5.1  The positions of the glycine substitutions and variants documented in this thesis.

The positions of the glycine substitutions (as indicated by arrows) are shown in relation to known substitutions in the whole α2(I) chain. In addition, the positions of the variants characterised are illustrated (indicated by triangles on the upper line). The map is a simplified version of that in figure 1.2; residue numbers have been omitted. Phenotypes are indicated below the glycine substitution positions. L, lethal; S, severe; Mo, moderate; O, osteoporosis; Va, variable.
patient who was only screened for mutations in the COL1A2 gene. Of the two patients with known mutations, one harboured a 9bp deletion in the \(\alpha(I)\) gene (Hawkins et al., 1991) and the other harbourd an \(\alpha(I)\)Gly988Cys mutation (Cohn et al., 1986). Using SSCP analysis, mutations were detected in the \(\alpha(I)\) gene in 10 of the patients (Mackay, 1992; Mackay et al., 1994) including the two known mutations which were used as controls. Eight mutations were similarly detected in the COL1A2 gene (one reported by Mackay, 1992 and Raghunath, Mackay, Dalgleish, Steinmann (1994), manuscript in preparation) and the others as reported in Chapters Four to Seven of this thesis). Excluding the patient for whom a mutation was not detected and who was only screened for a mutation in the COL1A2 gene, mutations in 17 patients out of 27 were identified. If all the patients are assumed to have a mutation within the region screened, this gives an overall efficiency of the SSCP detection method for OI mutations in the type I collagen genes as being 63.0%. A more accurate figure could have been calculated if it was known for certain that all the patients screened did indeed have mutations in either of the type I collagen genes, and if the mutations not identified had gone undetected by the SSCP method. This could be due to the mutation being completely resistant to detection by this method or not detected in any of the restriction digest combinations used in the analysis. If the nine patients for whom C1 peptide data were supplied (i.e. the patients who were expected to harbour a mutation within the region analysed), the efficiency of mutation detection becomes 100%. However, if only the mutations which agreed with the peptide data are considered this figure becomes 77.8%. Sheffield et al. (1993) reported a 79% detection rate for fragments below 212bp in length. Many of the fragments in this study were greater than this length and may have reduced the efficiency although shifts were seen in large (approximately 350bp) fragments.

The silent sequence change detected in patient 45 only manifested itself as one band shift. If this particular digest had not been used the sequence change would have gone undetected. This principle obviously also applies to OI mutations. Of the 28 patients, SSCP analysis failed to identify a mutation in 11. It is unlikely that all the remaining mutations are simply ones which escaped detection with this screening method. It is possible that some of the mutations may lie within regions of the mRNA that were not analysed. The SSCP analysis procedure was aimed at screening for defects that altered glycine codons; the whole of the triple helical domain was investigated thus other types of mutation may be the cause of the phenotype. Mutations that result in diminished gene expression of normal chains tend to result in a mild phenotype; it may be possible that some of the patients classified as having "non-lethal" OI fall into this category. The SSCP analysis approach for mutation detection would not have screened for defects in an allele that either (a) resulted in defective transcription ("null allele") or (b) allowed effective transcription but resulted in an unstable hnRNA which was subsequently degraded. Thus it would be necessary to determine the nature of the transcription step. However, investigation of the mRNA levels may still discriminate against the detection of a mutant allele because the screening is at the post-translational level.

It may be necessary, therefore, to screen at the gene level and to look for mutations that might result in defective gene processing. It would also have been ideal to have been able to carry out peptide analysis on the patient collagens to determine whether the mutation was within the triple helical region. Mutations at important intron donor and acceptor sites resulting in the skipping of exons should have been detected as a decrease in size of the PCR product. The smallest exon size in the region of the \(\alpha(2)\) mRNA screened is 45bp; if such an exon had been
skipped, the resulting PCR product would be 45bp smaller. Since all the PCR products were checked on polyacrylamide gels after each amplification, the difference in size of this product, as compared to the others, and the presence of a band of expected size from the normal allele in a heterozygote would have been noticeable. It is expected that SSCP would be more sensitive at detecting insertions and deletions than point sequence changes (Sheffield et al., 1993) presumably because they would have a more dramatic effect on the nature of the single stranded conformer. If any of the exons harbouring sequence complementary to that of a primer used in PCR was omitted from the mature cDNA, it is assumed that the amplification would fail due to the presence of only one operative primer. The absence of this exon would be confirmed from the reduced size of an overlapping PCR fragment since, for this reason, no two primers for use in SSCP analysis were designed from the same exon. If any of the mutations resulted in a deletion of a small number of bases, the difference in size would have been detected in the digestion fragments of the native DNA fragments on the SSCP gel even if it did not manifest itself as a band shift in the denatured digested samples. If the mutation was a single base deletion in a large fragment of one digest so that the size difference was not easily defined due to a lower degree of separation, it would be hoped that it would show up in a smaller fragment in another digest. Another possibility for the failure to detect all the mutations is that the proband may have a type of OI which is not linked to a type I collagen gene. However, only two instances have been documented in which the disorder may not be linked to either of the type I collagen genes (Aitchison et al., 1991; Wallis et al., 1993) and in such a small sample of patients it would seem unlikely that such patient would exist although the possibility cannot be ruled out.

9.4 MUTATIONS AT CpG DINUCLEOTIDES

Of the four different mutations presented in this thesis which resulted in OI, three occur at a CpG dinucleotide. The one exception is that which causes a glycine to glutamic acid substitution resulting from a mutation in the second position of a glycine codon (Chapter Seven). To investigate the frequency of occurrence of mutations at CpG dinucleotides, analysis of 65 point mutations, in both of the type I collagen genes, which resulted in an OI phenotype due to the substitution of a glycine residue was carried out (see section 9.6 for details). Of these, a mutation at a CpG dinucleotide which did not span an intron-exon boundary, was found in 14 patients. It has been suggested (Coulondre et al., 1978; Bird, 1980) that spontaneous base substitution hotspots may occur at 5-methylcytosine residues. The spontaneous deamination of the methylated base to thymine, which is not excised by the enzyme DNA-uracil glycosidase, may occur and unless correctly repaired, results in a G:C to A:T transition. This mechanism may be of major influence in the occurrence of glycine to serine and arginine mutations since a first position G to A transition in the GGN codon results in substitutions by a serine (AGC/AGT) or arginine (AGN/AGG) residue. The deamination of a methylcytosine on the antisense DNA strand in the complementary sequence of NNC-GGN-Xaa-Yaa would result in the first G of the glycine codon being replaced by an A. If the mutation took place on the sense strand the C in the third position of the preceding codon would be similarly altered resulting in a silent mutation with no OI phenotypic consequence.
Cooper and Youssoufflan (1988) analysed point mutations in a variety of human genetic diseases and reported that 35% occurred at CpG dinucleotides. It has been proposed that these hotspots could explain recurrent mutations in several genes (Youssoufflan et al., 1986; Wong et al., 1986; Abadie et al., 1989) including collagen (Pruchno et al., 1991). However, it has been suggested by Pruchno et al. (1991) that the COL1A1 gene may be protected from mutations arising as a result of deamination of methylcytosines at such dinucleotides by maintaining a low number of cytosines in the third position of a Yaa codon compared to average figures for the human genome. This serves to reduce the number of CpG dinucleotides occurring at the Yaa-Gly junctions. This protection system no doubt also occurs in the COL1A2 gene. The number of CpG dinucleotides which occur at a Yaa-Gly junction in the triple helical domain is 24. Of these, 23 occur at non-intron-exon boundaries. Thus 23, i.e. 7.8%, of all glycine codons in the COL1A2 gene could be mutated as a consequence of the deamination of a methylcytosine on the antisense strand of DNA, and give rise to an OI phenotype. Six such mutations have been identified to date.

All of the base changes resulting in an OI phenotype detected in this study were G to A transitions and the silent variants (excluding the PvuII RFLP from consideration) arose as a consequence of a C to T base change. With the exception of the base change in patients R and 45, all occurred at a CpG dinucleotide thus the OI phenotypes may have arisen by the deamination of a methylcytosine on the antisense strand of DNA and the silent variant in PE by an identical process on the sense strand. Of the mutations characterised by Mackay (1992) and Mackay et al. (1994), five of the OI-causing mutations were of a G to A nature — four occurring at a CpG dinucleotide — and one of the silent mutations was a G to A change at a CpG dinucleotide. The frequency of detection of these two types of sequence change in this study and that of Mackay (1992), suggests that these types of mutation were either detected more readily by SSCP analysis or occurred more often in the type I collagen genes than other types. Sheffield et al. (1993) carried out a study on various aspects of the SSCP detection system. In a study to test the effect of the nature of the base change on detection they detected G to A changes in 70% of cases whereas A to G, T to C and C to T changes were detected in 83%, 90% and 93% of cases respectively. These results tend to suggest that the high frequency of detection of G to A and C to T base changes in the SSCP studies is more likely to be attributable to the increased frequency of occurrence of such changes in type I collagen genes. In fact, during the analysis of 65 mutations (section 9.6), approximately 50% were seen to have occurred as a result of a G to A transition. The nature of the base substitution i.e. transversion versus transition, does not appear to be a major influence in the detection of a mutation by SSCP analysis (Sheffield et al., 1993).

9.5 USE OF ALTERNATIVE RADIOISOTOPES IN SSCP ANALYSIS

During the SSCP analysis of the patient samples, it was decided to investigate the use of $^{35}$S as an alternative radiisotope to $^{32}$P for incorporation into the amplification product. The main reason for this was the lower degree of hazard associated with the manipulation of $^{35}$S-labelled compounds as compared to those labelled with $^{32}$P. Both radionucleotides are $\beta$-emitters but the emission energy of $^{35}$S is approximately one tenth that of $^{32}$P (0.167 MeV maximum and 1.71 MeV maximum respectively). Initial experiments were carried out by simply substituting 0.5 µl (α-$^{32}$P)dCTP in the amplification reaction by 0.5 µl (α-$^{35}$S)dATP (section 2.14). The SSCP
\( \beta \)-fragment was chosen for the initial analysis and DNA from patient R (\( \alpha 2(I) \)Gly343Glu; Chapter Seven) was used as a control sample for a band shift. Amplification using \([\alpha-^{32}P]dCTP\) and the same patient samples as template was carried out in parallel. The products from \(^{35}S\) and \(^{32}P\) reactions were treated identically following amplification. Approximately 1/50 of the PCR reactions was checked on an agarose gel and autoradiographed overnight at room temperature (typical exposure conditions for an SSCP gel). The results are shown in figure 9.2a. As can be seen, the detection of \(^{32}P\) into the products was good but the signal from the \(^{35}S\) reactions was considerably lower. Although it seemed unlikely that the signal from the \(^{35}S\) reactions would be strong when the full length product was digested, 1 \( \mu \)l of each of the \(^{35}S\) and \(^{32}P\) samples was digested with the appropriate enzymes (Chapter Seven). One fifth reaction volume of each sample was electrophoresed through a 7% mini polyacrylamide gel which was autoradiographed at room temperature overnight. As expected, there was no signal from the \(^{35}S\) samples but a good signal from the \(^{32}P\) samples (results not shown).

Since the amplification step had been successful but displayed a low signal intensity, a larger amount of product was analysed in subsequent PCR reactions. A second PCR was carried out with \(^{32}P\) as before, however this time 2 \( \mu \)l of the product were digested with the appropriate enzymes. Denaturing solution was then added such that the resulting volume was smaller than normal but the final concentration was the same; 10 \( \mu l \) of a concentrated stock of denaturing solution was added to the 10 \( \mu l \) digestion volume. Denaturant was added in order to keep the conditions the same as for a regular SSCP sample: denaturation was not carried out. Two different volumes of the diluted digest samples, namely 5 \( \mu l \) and 15 \( \mu l \) (25% and 75% of the total volume respectively), were electrophoresed through a 7% mini polyacrylamide gel. The gel was autoradiographed for 60 hours at room temperature; results are shown in figure 9.2b. As can be seen, the digests went to completion and the signal was of moderate intensity. The overall resolution of the bands was better in the lanes where a smaller amount of the sample had been loaded. However, it was decided that these samples would not yield a good SSCP result since when DNA is denatured, the resulting single stranded fragments have a lower intensity than the double stranded fragment they were derived from and this would make the search for band shifts more problematic.

The final modification to the procedure was to double the volume of the radioactive PCR such that a 50 \( \mu l \) volume was amplified incorporating 1 \( \mu l \) of \([\alpha-^{35}S]dATP\). Following precipitation, the DNA was redissolved in 20 \( \mu l \) water and 4 \( \mu l \) of this solution were digested in a 20 \( \mu l \) volume. An equal volume of the concentrated denaturing solution was added following incubation. A digestion check gel was run and the autoradiograph indicated that all the reactions had gone to completion and that the signal was of a fairly high intensity. Hence the samples were electrophoresed through an MDE SSCP gel. Each sample loaded consisted of 5 \( \mu l \) of digested DNA and 5 \( \mu l \) dye. An overnight exposure of the gel yielded visible, sharp bands. However, due to the amount of DNA that had to be loaded onto the gel to be sure of a high intensity, much of it had failed to denature. This combination of double and single stranded fragments was too confusing for the interpretation of the gel and thus the conclusion from this series of experiments was that \(^{35}S\) is not a suitable nuclide for use with the SSCP mutation detection method.

The main reason for the non-suitability of \([\alpha-^{35}S]dATP\) in SSCP analysis is most likely to be due to the fact that \(^{35}S\) has a lower energy than \(^{32}P\) and thus results in a fainter signal. In addition to the lower energy, it is
Figure 9.2 Comparison of $^{35}$S- and $^{32}$P- incorporated PCR amplification products

(a) Four patient DNA samples were amplified using primers A2A and SSCP1A2βR incorporating radioactive nucleotide. The results for the incorporation of $^{32}$P and $^{35}$S are presented. It can be seen that the incorporation of [$\alpha$-$^{35}$S]dATP is inferior to that of [$\alpha$-$^{32}$P]dCTP for each sample.

(b) Digestion products derived from [$\alpha$-$^{35}$S]dATP PCR amplifications are presented to illustrate the large quantity required of such a radioactive sample to yield a visible banding pattern on a polyacrylamide gel. 5 μl represent 25% reaction volume.

Fragment sizes are not given since they are not relevant.

Figure 9.3 Comparison of $^{33}$P- and $^{32}$P- incorporated PCR amplification products

(a) Amplified β-fragment products incorporating [$\alpha$-$^{33}$P]dCTP are presented for the four patients. The incorporation levels can be compared to those in figure 9.2(a).

(b) Digestion products of the $^{33}$P-incorporated amplification products are shown. The varying patterns seen in the AluI/Ddel lanes is due to the presence of the FvαII RFLP. The intensity and sharpness of the signals is to be observed.

Fragment sizes are not given since they are not relevant.
possible that the structure of the $^{32}$S-dATP may result in poorer incorporation during amplification. Whereas $^{32}$P-dCTP is a natural base with the α-phosphorus radiolabelled, $^{35}$S-dATP is an analogue of dATP. Because of this, during the amplification process, Taq DNA polymerase may exclude $^{35}$S-dATP from the product in favour of the incorporation of the natural dATP base. With these ideas in mind, the use of another radionuclide was investigated, namely $^{33}$P-dATP. The $^{33}$P nuclide has a $T_{1/2}$ of 25 days and an emission energy of 0.249 MeV which is approximately one seventh that of $^{32}$P and double that of $^{35}$S. Thus it is still much safer to work with than $^{32}$P-labelled nucleotides.

PCR amplification was carried out using 0.5 μl of $^{33}$P-dATP per 25 μl reaction volume. The patient samples and primer pair that were used in the $^{33}$P/$^{35}$P isotope comparison were used again. Approximately 1/50 of the amplification products was analysed on a 0.8% agarose gel which was autoradiographed overnight at room temperature. Figure 9.3a illustrates that the signal obtained, and thus the incorporation of radio-labelled nucleotide, was good. Since the signal was not as strong as that obtained through using $^{32}$P, 2 μl of each reaction were used in the digestion reactions. One third of each digest volume was electrophoresed through a 7% mini polyacrylamide gel which was autoradiographed overnight at room temperature. As can be seen from figure 9.3b the bands were distinct and of sufficient intensity for interpretation. Since the results appeared promising, samples were electrophoresed through an MDE SSCP gel. Each lane was loaded with 3.5 μl of digested DNA and 3.5 μl of dye. The gel was autoradiographed at room temperature. A readable signal was obtained after a 5 hour exposure and a 48 hour exposure revealed an informative pattern in which the band shifts expected in the denatured 'R' DNA sample were clearly visible (results not shown).

The conclusions from this series of SSCP experiments are that $^{32}$P-dATP is not a suitable alternative radio-labelled nucleotide for incorporation into a PCR amplification product whereas $^{33}$P-dATP is. At the time this study was carried out, the $^{33}$P-labelled nucleotide was considerably more expensive than $^{32}$P-dCTP and was thus considered too expensive for routine screening. However, if in future the demand for the nucleotide is such that it could be offered at a more competitive price, it would be ideal for use in similar analyses since it is less hazardous to work with than $^{32}$P-labelled nucleotides. Ideally, a non-radioisotopic method would be used for routine screening. However, the size of the gel and its fragility are such that staining methods might be difficult although some have been reported (section 1.6.2.1 and a recent modification by Chaubert et al., 1993). Such methods also require a larger amount of DNA to be loaded on the gel; this might result in reduced denaturation of samples.

9.6 A VARIANT SITE SCREENING STRATEGY FOR MUTATIONS CAUSING OSTEOGENESIS IMPERFECTA

9.6.1 Introduction

It had been observed that the mutations characterised in this study all resulted in the alteration of restriction enzyme recognition sequences. The G to A base change in patients PI/KO, DW/23/PE and SC destroyed a BsuUI site and the G to A base change in patient R destroyed an AvrII site. It was thought that many mutations
may result in similar enzyme site changes and that this concept could be exploited and used as a preliminary screening strategy for patients with suspected OI mutations; the presence of an RFLP would indicate within a few bases where the potential mutation occurs. In order to determine the most appropriate enzymes to use, the analysis of sequence surrounding a number of known mutations in both COL1A1 and COL1A2 was carried out. The majority of the mutations which were analysed were the single base changes resulting in a glycine substitution documented in the review by Kuivaniemi et al. (1991). Each of these mutations was localised on the cDNA sequences of COL1A1 and COL1A2 and 10 bases of sequence either side of the mutated base were recorded. A total of 56 mutations were studied in this manner along with the mutations characterised in this and a previous study from our laboratory (Maclcay, 1992). Manual analysis of these data indicated that many of the mutations would result in a restriction endonuclease site change. The ones that appeared to be the most useful were recorded. The results are summarised in table 9.2.

Of the five enzymes described, a BstUI site was destroyed by only three different mutations out of a possible 65, i.e. 4.6%, and was therefore not considered to be a suitable enzyme for the proposed general screening purpose. An ScrFI site, on the other hand, was destroyed by 24 different mutations out of the 65, i.e. 36.9%, but due to the fact that the recognition site of this enzyme is 5' CCNGG 3', a sequence common in collagen genes, the fragments generated by a digest with this enzyme would be numerous and of small size, thus it would be difficult to detect any variant sites present due to poor separation on a gel. Hence, these two enzymes were eliminated from the experiment and the figures in table 9.2. were adjusted accordingly: whereas before, using all five enzymes, 65% of all a1(I) mutations and 77% of all a2(I) mutations could, in theory, be detected, eliminating BstUI and ScrFI reduced these figures to 55.8% and 54.5% respectively — an average of 55.2% for type I collagen point mutations resulting in a glycine codon change. Thus it appeared that this strategy could potentially detect a large number of glycine codon mutations. Patients for whom a negative SSCP result had been obtained and those in which a positive SSCP result had lead only to the characterisation of a sequence variant, were chosen for screening using this RFLP approach.

### 9.6.2 Variant site screening strategy

The four primer pairs from the SSCP screening strategy were employed to amplify overlapping regions of the α2(I) gene (see section 3.3.1). Amplifications of all four regions were performed on the cDNA from the patients. It was noted that although the Avall and HaeIII digests of each amplification product would give rise to a series of fragments within a similar size range, the MspI digest would yield fragments too large to be analysed on the same percentage polyacrylamide gel as those of Avall and HaeIII for the β and δ regions respectively, thus for these amplification products a double digest was carried out namely MspHIinfl, and MspIIinfl for the β and δ regions respectively. Fragment sizes for all three restriction digests of all four amplification products are given in table 9.3.

The α2(I) cDNA was amplified as in the SSCP protocol substituting [α-32P]dATP for [α-32P]dCTP since this gave a sharper signal with more clearly defined bands (see section 9.4). Products were isopropanol precipitated and 2 µl were incubated in a 10 µl single or double restriction enzyme digest reaction for 1.5 hours.
<table>
<thead>
<tr>
<th>TYPE I COLLAGEN CHAIN</th>
<th>NUMER OF MUTATIONS DETECTABLE BY COMBINATIONS OF RESTRICTION ENDONUCLEASES</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AvuII + BsuUI + MspI + ScrFI</td>
<td>AvuII + BarUI + MspI + ScrFI</td>
</tr>
<tr>
<td>α1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>α2</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>11</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 9.2  Restriction endonuclease sites destroyed by known OI mutations

The number of times that an OI mutation caused the disruption of a restriction endonuclease recognition site is presented. The recognition sites of the five enzymes for potential use in analysis are indicated. The total number of different mutations detectable by these restriction enzymes is indicated in bold. When a mutation can be detected by a combination of the enzymes, the number is recorded in the latter columns.
<table>
<thead>
<tr>
<th>α2(1) SSCP FRAGMENT</th>
<th>ENZYMES EMPLOYED</th>
<th>FRAGMENTS GENERATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>AveII</td>
<td>243, 90, 81 (x 2), 75, 72, 58, 54, 45, 36, 18 (x 3), 12, 9 (x 5)</td>
</tr>
<tr>
<td></td>
<td>HaellI</td>
<td>279, 111, 81 (x 2), 72, 60, 36, 33, 27 (x 6), 22, 9</td>
</tr>
<tr>
<td></td>
<td>MspII</td>
<td>292, 261, 126, 81, 45, 39, 36, 27, 18, 15, 6</td>
</tr>
<tr>
<td>β</td>
<td>AveII</td>
<td>117, 97, 90, 81, 72, 63 (x 4), 54, 45, 36 (x 2), 27 (x 2), 18 (x 2), 9 (x 7), 6</td>
</tr>
<tr>
<td></td>
<td>HaellI</td>
<td>180, 172, 158, 109, 99, 81, 54, 44, 27 (x 3), 25, 22, 14</td>
</tr>
<tr>
<td></td>
<td>MspII/HindII</td>
<td>321, 189, 120 (x 2), 87, 81 (x 2), 40</td>
</tr>
<tr>
<td>γ</td>
<td>AveII</td>
<td>185, 99, 81 (x 2), 72, 63 (x 3), 36 (x 3), 30, 27 (x 5), 22, 18 (x 3), 9 (x 9), 4</td>
</tr>
<tr>
<td></td>
<td>HaellI</td>
<td>314, 216, 187, 73, 72 (x 2), 54, 51, 38, 35, 18, 11</td>
</tr>
<tr>
<td></td>
<td>MspI</td>
<td>387, 277, 236, 126, 115, 2</td>
</tr>
<tr>
<td>δ</td>
<td>AveII</td>
<td>153, 93, 81, 63, 54 (x 2), 36, 31, 27 (x 2), 18 (x 3), 9 (x 3)</td>
</tr>
<tr>
<td></td>
<td>HaellI</td>
<td>187, 116, 101, 81, 63, 46, 18 (x 4), 14, 11, 9</td>
</tr>
<tr>
<td></td>
<td>MspII/RsaI</td>
<td>207, 184, 172, 126, 11</td>
</tr>
</tbody>
</table>

Table 9.3  
Restriction digests for variant site analysis of the α2(1) mRNA

Details of restriction digests and resulting fragments are indicated for each of the α2(1) SSCP PCR fragments. The fragments are listed in order of descending size.
Upon completion of digestion, 5 μl loading dye (section 2.12.1) were added and 4 μl of this diluted mix were loaded onto a sequencing gel-sized 8% polyacrylamide gel cast as in section 2.12.3. Eight samples of the same digest were run side by side and all three digests of one amplification product were run on one gel. The samples were electrophoresed at room temperature at 5 W for 5 hours alongside radioactively end-labelled pUC13/HpaII digestion fragments (section 2.6) as size markers.

9.6.3 Results

A sample with a known mutation was included on the first gel. This was the G to A transition in SSCP amplification fragment β which resulted in the substitution of glycine-343 by glutamic acid in the mature α2(I) chain. This mutation is documented in Chapter Seven and results in the destruction of an AvaII site. The results from the first gel showed that this RFLP could be detected readily (results not shown), however, no variants in the other DNA samples could be seen in the β-fragment or when the analysis was applied to the remaining three fragments (α, γ and δ). The analysis was repeated on the same patient samples to look for mutations in the COL1A1 gene which may have gone undetected in the initial SSCP screening (Mackay, 1992); the same four primer pairs as used in the study were utilised in this project. The same enzymes as for the RFLP screening of COL1A2 were utilised in the digestion of the four fragments generated by these COL1A1 primer pairs. The details of the primers and fragment sizes for the digests performed are given in tables 9.4 and 9.5 respectively. Some of the parameters differ from those quoted by Mackay (1992). This is probably a consequence of minor differences in the calculation of primer concentration, buffer composition etc. that necessitated the recalculation of optimal conditions for PCR amplification. Controls were included when available but only one resulted in a visible change in a band. An MspI RFLP in one patient (Mackay, 1992) was detected as a reduced intensity of two smaller fragments. The new fragment is only 1bp different in length to an existing 225bp fragment and the two could not be distinguished (results not shown). Again, no novel mutations were identified.

9.6.4 Discussion

There are several reasons why no novel mutations were detected by this approach despite the high number of reported mutations which result in RFLPs:

(i) The mutation may result in two small fragments giving rise to a slightly larger one which is of equal size to a normal digestion product. With small products, differences in intensity may be slight and thus this type of mutation would be difficult to detect. This was highlighted by the fact that a cDNA sample with a known mutation was analysed alongside unknown samples. In this case, an AvaII RFLP resulted in a 45bp fragment from adjacent 36bp and 9bp fragments and the RFLP was not detected due to the existence of normal fragments of this size.

(ii) The AvaII digests of all four amplification products generated many fragments and often two or three of a given size. It is possible that if a mutation resulted in the disappearance of two bands which are each just one of two or three bands of an identical size, to give rise to a band of equal size to yet another band, the subtle change of intensities may again be difficult to detect.
<table>
<thead>
<tr>
<th>PRIMER IDENTIFICATION</th>
<th>α1(I) SSCP FRAGMENT AMPLIFIED</th>
<th>SIZE OF PRODUCT bp</th>
<th>PRIMER SEQUENCE 5' → 3'</th>
<th>POSITION OF PRIMER</th>
<th>AMPLIFICATION CONDITIONS</th>
<th>NUMBER OF CYCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSCP1A1αL</td>
<td>α</td>
<td>1210</td>
<td>CCACGCTCCTGCTCCCCTTTAG</td>
<td>Exon 1</td>
<td>95°C – 1.4 min</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCCAGGAGCCACCAAGAATAC</td>
<td>Exon 19</td>
<td>55°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.4 min</td>
<td></td>
</tr>
<tr>
<td>SSCP1A1βL</td>
<td>β</td>
<td>1054</td>
<td>GCTGATGGAAGCCCTGAGTGC</td>
<td>Exon 18</td>
<td>95°C – 1.4 min</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCCCTGGAAGACCCAGGCTGCA</td>
<td>Exon 32</td>
<td>55°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.4 min</td>
<td></td>
</tr>
<tr>
<td>SSCP1A1γL</td>
<td>γ</td>
<td>926</td>
<td>CCGGCGCTTACCAGGAATGCC</td>
<td>Exon 32</td>
<td>95°C – 1.4 min</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GCGCCAGGAGAAGGCTTCCG</td>
<td>Exon 42</td>
<td>55°C – 1.0 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>70°C – 1.2 min</td>
<td></td>
</tr>
<tr>
<td>COL1A1L</td>
<td>δ</td>
<td>936</td>
<td>CCGAAGGTTTCCCCCTGGACGA</td>
<td>Exon 43</td>
<td>95°C – 1.4 min</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CGGTCATGCTCGGAGACGAC</td>
<td>Exon 50</td>
<td>70°C – 1.2 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>72°C – 1.4 min</td>
<td></td>
</tr>
</tbody>
</table>

Table 9.4 Primers used for the variant site analysis of the α1(I) mRNA

Details of the oligonucleotides used to generate four overlapping fragments along the length of the α1(I) collagen mRNA for variant site analysis are presented. Conditions for amplification are given and may differ from those in the original application (Mackay, 1992).
<table>
<thead>
<tr>
<th>α1(I) SSCP</th>
<th>ENZYMES EMPLOYED</th>
<th>FRAGMENTS GENERATED</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td><strong>Avall</strong></td>
<td>314, 171, 153, 137, 93, 54, 45, 36 (×3), 27, 18 (×2), 9 (×8)</td>
</tr>
<tr>
<td></td>
<td><strong>HaeIII</strong></td>
<td>222, 163, 126, 82, 75, 72, 64, 46, 44, 42, 36 (×3), 35, 23, 22, 18 (×2), 17, 12 (×2), 9 (×5)</td>
</tr>
<tr>
<td></td>
<td><strong>MspI</strong></td>
<td>252, 225, 220, 154, 99, 78, 71, 66, 18, 9 (×3)</td>
</tr>
<tr>
<td>β</td>
<td><strong>Avall</strong></td>
<td>138, 117, 90 (×2), 88, 63, 54 (×3), 42, 36 (×2), 27 (×3), 21, 18 (×3), 9 (×4)</td>
</tr>
<tr>
<td></td>
<td><strong>HaeIII</strong></td>
<td>162, 108 (×3), 83, 81, 80, 79, 54, 47, 45, 36, 27, 18, 9 (×2)</td>
</tr>
<tr>
<td></td>
<td><strong>MspI</strong></td>
<td>189, 156, 144, 117, 108, 101, 81, 68, 45, 27, 18</td>
</tr>
<tr>
<td>γ</td>
<td><strong>Avall</strong></td>
<td>135, 99, 76 (×2), 63 (×2), 45 (×5), 27 (×4), 18 (×3), 9 (×3)</td>
</tr>
<tr>
<td></td>
<td><strong>HaeIII</strong></td>
<td>93, 90, 80, 72, 64 (×2), 63, 54, 46, 45 (×3), 35, 27, 26, 18 (×4), 5</td>
</tr>
<tr>
<td></td>
<td><strong>MspI</strong></td>
<td>315, 248, 129, 106, 99, 27</td>
</tr>
<tr>
<td>δ</td>
<td><strong>Avall</strong></td>
<td>426, 90, 81, 72, 69, 36 (×2), 27, 18 (×3), 9</td>
</tr>
<tr>
<td></td>
<td><strong>HaeIII</strong></td>
<td>284, 117, 96, 72 (×2), 63, 61, 53, 45, 18 (×3), 10, 9</td>
</tr>
<tr>
<td></td>
<td><strong>MspI</strong></td>
<td>279, 259, 144, 109, 87, 58</td>
</tr>
</tbody>
</table>

Table 9.5  Restriction digests for variant site analysis of α1(I) mRNA.

Details of restriction digests and resulting fragments are presented for each of the α1(I) SSCP PCR fragments. The fragments are listed in order of descending size.
(iii) If two fragments from bands containing more than one fragment gave rise to a band of such a length that it did not separate out efficiently from the uppermost fragments on the gel, the change in intensities and the presence of the large band may be difficult to determine. This could be overcome by electrophoresing the samples on a lower percentage acrylamide gel to separate out the larger digest fragments.

(iv) The mutation may lie outside the boundaries of the regions covered by this analysis.

Differences in the \textit{HaeIII} and \textit{MspI} digests should be more easy to detect due to the smaller number of fragments generated in the digest but these enzymes could potentially detect fewer known mutations than \textit{AvaII}.

In order to try and overcome these problems, preliminary experiments were carried out based on the method of Uitterlinden \textit{et al.} (1989). The approach undertaken was to digest patient PCR amplified cDNA and to electrophorese the fragments through a non-denaturing polyacrylamide tube gel. Once these had run, the gel could be transferred to a denaturing polyacrylamide gel and electrophoresed in the second dimension. Under such conditions, the multiple bands of identical size should be separated out according to their sequence composition. If a sample from a normal individual was electrophoresed alongside that of the expected variant, any discrepancies between fragment positions would localise a variant to a particular digest fragment. This could then be investigated further by DNA sequence analysis. Unfortunately only preliminary experiments could be carried out in the time available (Rose and Dalgleish, unpublished data) but it would be interesting to see if this approach detected the variants expected from the above method.

9.7 FUTURE WORK

In order to determine the underlying mutation causing the OI phenotype in the other patients in this study it will be necessary to use a different system. There are many different detection systems published. One is that of Ganguly \textit{et al.} (1993) which claims to be able to detect all point mutations and was developed while searching for mutations in collagen genes. The method relies on the separation of heteroduplexes on an acrylamide gel containing a Tris-taurine buffer, formamide and ethylene glycol. This system is being investigated within this laboratory using the patients from this study for which mutations were not characterised (Gillies, Rose and Dalgleish, unpublished data). Preliminary results indicated that a known mutation (patient PE, ct\(\text{G}\)502S; Chapter Five) could be detected as a shift (results not shown). Another approach is dideoxy fingerprinting. This method is a mixture of dideoxy sequencing and SSCP analysis that can detect the presence of single base changes in PCR-amplified DNA. Sarkar \textit{et al.} (1992b) detected 100\% of mutations using this method. A Sanger sequencing reaction is carried out with one dideoxynucleotide and the products are resolved on a non-denaturing polyacrylamide gel to give a ladder of fragments. Three patterns were seen which indicated the presence of a mutation. These were the absence of a band at the site of mutation and a shift of some or all bands beyond the point of mutation.

These are just two alternative methods that could be employed and although they may be quoted as having a potential 100\% detection success, they may not detect all of the remaining sequence defects.
CHAPTER TEN

INVESTIGATION OF ALTERNATIVE SOURCES OF COLLAGEN GENE TRANSCRIPTS TO AID MUTATION ANALYSIS

10.1 INTRODUCTION

The regulation and expression of housekeeping and tissue-specific genes differ. The former are ubiquitously expressed in order to produce the proteins which are required by all cell types for basic functions whereas the latter are developmentally expressed and transcripts can only be detected in the tissues which express the protein. However, it has recently been shown that a low level ubiquitous transcription exists, probably for all different mRNA species. This was introduced in Chapter One. The phenomenon was demonstrated by Chelly et al. (1988) who reported the detection of dystrophin gene transcripts in both muscle cells which possess the protein, as well as non-muscle cells which would not be expected to harbour the protein. They found that using the cDNA PCR amplification method, products from muscle tissues such as skeletal, heart and smooth muscle could be detected by simple ethidium bromide staining following gel electrophoresis. However, in the non-muscle tissues that were investigated (both foetal, adult and cultured cells of non-muscle origin) the amplified material could only be detected following probe hybridisation. For the lymphoblastoid RNA material, these products were shown to be those expected by restriction endonuclease digestion and analysis of resulting fragments. One control experiment involved the amplification of mRNA derived from a patient who suffered from Duchenne muscular dystrophy (DMD) and possessed a 4Mb DNA deletion which spanned the entire DMD gene. In this case, no product was produced from lymphoblastoid total mRNA. The number of dystrophin transcripts in these non-muscle cells was estimated to be one copy per 500–1000 cells. This may represent a 100-fold decrease in frequency when compared to 'low abundance' mRNA sequences and is not thought to have any physiological significance (Chelly et al., 1988). Such low level ubiquitous expression was detected for other genes (Chelly et al., 1988, 1989) and was termed basal transcription by the authors. This phenomenon was independently recognised by Sarkar and Sommer (1989) and termed illegitimate transcription and has been detected for other transcripts in a variety of tissue sources. These transcripts have been reviewed by Kaplan et al. (1992) and include factor VIII mRNA transcripts from lymphocytes (Berg et al., 1990), types I, II and III collagen genes from fibroblast and lymphoblastoid cells (Chan and Cole, 1991), spermatid specific transcripts from peripheral blood lymphocytes in both males and females (Stonkski et al., 1991), the COL4A5 gene from lymphoblast cells (Kuehlemann et al., 1992) and the CFTR gene from lymphocytes and lymphoblastoid cells (Fouknechten et al., 1992). Handt et al. (1992) also witnessed ectopic transcription of the parathyroid hormone gene in a number of human, rat and bovine tissues, namely lymphocytes, lymphoblastoid cells and a rat tumour. The term ectopic has also been used to describe transcription of this nature.

The products of illegitimate transcripts have been shown to mirror the sequence of the mature mRNA for the product which has been isolated from a tissue harbouring the protein (Chelly et al., 1991a, b). This indicates
that the transcripts are not incorrectly spliced, derived from an incorrect initiation event or represent nuclear RNA precursors (Kaplan et al., 1992). The consequences of such a phenomenon for the study of gene mutation is evident. Since the transcripts are the same as legitimate transcripts, they can be used in the study of genes in disease. Indeed, their existence could resolve the problems inherent in the analysis of diseases where the gene transcripts may be inaccessible or have a specific expression pattern. It has been demonstrated that DMD and BMD (Becker muscular dystrophy) carrier females can be identified by the amplification of the entire coding region of the dystrophin mRNA from lymphocytes (Roberts et al., 1990).

The aim of this particular study was to identify a source of collagen transcripts based on the basal transcription phenomenon. If successful, it was hoped that this source could be used in two applications, namely the detection of mutations causing collagen disorders such as OI and EDS, and the production of a library enriched with collagen sequences. The latter is discussed in detail in Chapter Eleven. The current ways in which collagen gene defects are identified are focused, in the main, on mRNA derived from cultured fibroblast cells. Since the major collagen disorders result from defects in fibrillar collagen genes (Chapter One) fibroblasts derived from skin punch biopsies are suitable sources for the transcripts. However, for other disorders which are the result of defects in less abundant collagens, transcripts may not be accessible from this source. An example is the Schmid metaphyseal chondrodysplasia which has been found to result from a defect in the COL10A1 gene (Warman et al., 1993); the expression of type X collagen is restricted to hypertrophic chondrocytes. An alternative source would be an advantage in the detection of the gene and possible diagnosis. If a source of collagen transcripts could be identified in which no collagen species predominated, theoretically, any mutation in any gene could be identified and such a source might be important in mutation characterisation and subsequent familial diagnosis.

For this purpose, blood was proposed as the substrate of choice. Vaage and Lindblad (1990) showed that under identical culture conditions, murine peritoneal macrophages produced as much type I collagen as mouse tail fibroblasts. Macrophages mature from monocytes and either circulate in the blood stream or move from the blood and become involved in phagocytosis in response to foreign material in tissues. Since it is known that macrophages produce type I collagen (Vaage and Lindblad, 1990; Vaage and Harlos, 1991) it is probable that monocytes also have a latent capacity to synthesise the protein. It is possible that the transcripts may be at an elevated level compared to basal levels but not overabundant as in skin fibroblasts. From this it can be proposed that blood, which contains monocytes, is a potential good source of collagen transcripts.

Blood is a highly accessible bodily material. In order to identify defective genes in patients with OI or EDS, it is usual to culture fibroblasts derived from skin. However, this is a more invasive approach than taking blood and for patients with the milder forms of OI or osteoporosis, the removal of blood may be more acceptable. In addition, skin punch biopsies need to be cultured prior to the harvesting of mRNA and/or genomic DNA for subsequent analysis whereas blood can be treated immediately to extract nucleic acid. On the other hand, the advantage of culturing fibroblasts is that the characteristics of the collagen proteins can be investigated if so required. In the detection of mutations, it has been shown in previous chapters that protein data are not always required although in the prediction of the mechanics of defective proteins they are more important. Thus in this study, the amplification of sequences from collagen transcripts derived from total blood was investigated.
10.2 THE APPROACH FOR THE STUDY

A long-term strategy was proposed for the study. It was deemed necessary to ensure that each of the stages would work accurately and efficiently before advancing to the next. This approach is outlined below:

(i) Choice of transcript for investigation.

It was decided that two or more short stretches of DNA should be analysed in the first instance. If the strategy worked for short fragments, it could later be applied to longer fragments such as those involved in SSCP mutation detection analysis. One of these transcripts was to be type III collagen since there were two potentially informative RFLPs in close proximity to each other.

(ii) Confirm that frozen blood can be used for the production of total RNA as well as fresh blood.

This was thought to be necessary since one of the aims of the project was to find a source of basal level transcripts to use for the detection of mutations in heritable disorders. It could not be assumed that a patient will be available to give a blood sample immediately prior to nucleic acid preparation — the blood is more likely to be stored in citrate tubes at -20°C.

(iii) Confirm that amplification of sequences from blood-derived RNA is possible.

If the PCR amplification of sequences was difficult or even impossible, another source of transcripts would have to be investigated.

(iv) Confirm that products from RNA derived from blood are the same as those generated from fibroblast RNA.

Since RNA isolated from cultured skin fibroblasts were freely available in the laboratory, amplification products from this RNA were considered to be a suitable positive control for ectopic transcript amplification. Both type I and III collagen have been analysed at the cDNA level via total RNA isolated from these cultured cells. The analyses have resulted in the identification of many mutations underlying phenotypes in both OI and EDS patients (see this study; Mackay, 1992; Mackay, Steinman, De Paepe, Dalgleish, manuscripts in preparation) and sequence discrepancies, other than rare variants or single base Tαq DNA polymerase-induced errors, were not encountered.

(v) Compare haplotypes at the type III collagen RFLPs in genomic and complementary DNA.

The assumption made in this study was that if the true transcripts were being amplified, both genomic and complementary DNA from the blood of a single individual would yield the same haplotypes at both polymorphic loci. This was thought to be a suitable control experiment for proving the correct, or otherwise, identity of the amplified transcript.

(vi) Apply the technique to known and unknown OI mutations.

If the results were encouraging, it was hoped that blood samples from the patients with non-lethal OI described in Chapters Four and Six could be obtained from the collaborating laboratory. These would constitute
positive controls for the detection of mutations underlying OI using the SSCP analysis. If this was successful, the approach could be applied to patients with uncharacterised mutations. Obviously, unless a blood sample had been taken at birth, collagen gene mutations in probands with lethal OI could not be characterised by such an approach. However, if the mutation in the affected individual was identified by conventional approaches, the possibility of parental somatic mosaicism for the mutation could be investigated via blood samples.

10.3 RESULTS

The results will be introduced under the proposed areas of progressive study.

(i) **Choice of transcript for investigation.**

In order to investigate the feasibility of such a product, it was decided initially to amplify two different short stretches of cDNA derived from total blood mRNA to see if products of the expected size were generated. The sequences chosen were from COL1A1 and COL3A1 cDNA. Total mRNA was isolated from fresh blood as in section 2.10.2 and first strand cDNA synthesis was carried out as in section 2.11.3. Standard PCR amplification was then effected. Details of the two primer pairs (COL1A1-L/COL1A1-R and COL3A1-L/COL3A1-R), the regions amplified and PCR conditions are indicated in table 10.1. The region flanked by the type I collagen primers harbours two sequence polymorphisms that destroy single Alul and HaeII restriction endonuclease recognition sequences (see figure 10.1). These RFLPs were reported by ZafaruUah et al. (1990) and Tromp et al. (1991) respectively. Primers for the amplification of α1(I) sequences were available in the laboratory (Hawldns, 1991). Those for the type III collagen sequences were designed from sequence reported by Ala-Kokko et al. (1989).

(ii) **Confirm that frozen blood can be used for the production of total RNA as well as fresh blood.**

A personal blood sample had been kept at -20°C for a period of 12 months in a citrate tube. A fresh sample of my blood was taken and transferred to an identical tube. Total RNA was prepared from each sample in parallel. Following RNA isolation, aliquots were analysed on an agarose gel. The isolates did not differ widely in either yield or quality (data not shown), thus it was assumed that RNA from both fresh and frozen blood sources were adequate for the analysis proposed.

(iii) **Confirm that amplification of sequences from blood-derived RNA is possible.**

It was unlikely that genomic DNA contaminated the mRNA prepared from blood however, in order to eliminate this possibility, the primers were designed such that they resided in different exons. It was expected that this would allow genomic-derived amplification products to be distinguished from cDNA-derived ones due to the large size of collagen gene introns. The amplification of fibroblast cDNA was carried out in parallel with that of blood cDNA. Preliminary results indicated that the products obtained from each of the primer pairs was of the expected size when compared to those obtained from fibroblast cDNA. Contaminating genomic fragments were not
<table>
<thead>
<tr>
<th>PRIMER IDENTIFICATION</th>
<th>COLLAGEN TYPE</th>
<th>SIZE OF PRODUCT bp</th>
<th>PRIMER SEQUENCE 5'→3'</th>
<th>POSITION OF PRIMER</th>
<th>AMPLIFICATION CONDITIONS</th>
<th>NUMBER OF CYCLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL1A1-L 1</td>
<td>α(I)</td>
<td>250</td>
<td>CGGAAGGTTCCTGGACGA</td>
<td>Exon 42</td>
<td>95°C - 1.4 min</td>
<td>30</td>
</tr>
<tr>
<td>COL1A1-R 1</td>
<td></td>
<td></td>
<td>GGCCTGTCGCCCTGTCTCA</td>
<td>Exon 45</td>
<td>64°C - 1.2 min, 70°C - 1.0 min</td>
<td></td>
</tr>
<tr>
<td>COL3A1-L</td>
<td>α(Ill)</td>
<td>224</td>
<td>GAACGTGACCTCTGATTT</td>
<td>Exon 31</td>
<td>95°C - 1.0 min</td>
<td>35</td>
</tr>
<tr>
<td>COL3A1-R</td>
<td></td>
<td></td>
<td>CCTGGGACACCATCAGCACC</td>
<td>Exon 33</td>
<td>60°C - 1.0 min, 70°C - 1.0 min</td>
<td></td>
</tr>
<tr>
<td>COL3FL</td>
<td>α(Ill)</td>
<td>675</td>
<td>TCCACAAGGATTACAAGGCT</td>
<td>Exon 28</td>
<td>95°C - 1.4 min</td>
<td>35</td>
</tr>
<tr>
<td>COL3FR</td>
<td></td>
<td></td>
<td>CTCTTGACACCTTGGGACC</td>
<td>Exon 38</td>
<td>59°C - 1.0 min, 70°C - 1.4 min</td>
<td></td>
</tr>
</tbody>
</table>

Table 10.1 Primers used for the amplification of basal level transcripts

Details of the oligonucleotides used to amplify sequences from α(I) and α(Ill) cDNA for the purpose of basal transcription analysis are presented. Conditions given for amplification using the α(I) primers differ from those quoted in the original application. 1 Hawkins et al. (1991).
Figure 10.1  The region of type III collagen cDNA investigated in the basal transcription study

The region of type III collagen cDNA investigated in the basal transcription study and surrounding sequence is indicated and is drawn to scale. The bold line represents the collagen sequence with relevant restriction enzyme sites indicated. Sa, SaeI (SrfI); H, HaeIII; A, Alul; P, PvuII. The polymorphic sites are indicated with asterisks. The positions of the nested primer pair is given in (a) and of the flanking primer pair in (b). The cDNA probe used in the hybridisation analysis (SS-819) is illustrated in (c).
encountered. It was decided to investigate the \(\alpha(I)\) product in detail. The reason for investigating this in preference to the \(\alpha(I)\) PCR product was due to the presence of the two easily identifiable RFLPs within the fragment. In addition, the primer pair used to generate this fragment was newly synthesised and as such had not been used in the laboratory before. Therefore, no identical fragments could be capable of contaminating any initial experiments. The primers for the \(\alpha(I)\) fragment had been used in a previous application (Hawkins et al., 1991).

(iv) **Confirm that products from RNA derived from blood are the same as those generated from fibroblast RNA.**

First strand cDNA synthesis was carried out on RNA from frozen blood (FRO), fresh blood (FRE) and from a randomly selected cultured fibroblast cell line (CL). A nested primer PCR approach was adopted in order to optimise primer annealing specificity. Initial amplification was carried out using the primers COL3FL and COL3FR and a fraction of the product was used to seed a subsequent amplification employing the primers, COL3A1-L and COL3A1-R (table 10.1). Following the nested primer amplification, a band of the expected size (224bp) was observed in each case. The greatest yield was obtained in the CL RNA sample (results not shown).

Initial trial digestions of the CL product with \(Alul\) and \(HaeIII\) had shown that the reaction did not always go to completion even when the PCR products were purified as in section 2.2.3. Spermidine trichloride was included in the digestion reactions since this has been suggested to be an enhancer of enzyme activity (Pingoud, 1985) but no dramatic improvement was seen. However, the addition of casein (section 2.2.2) to the reactions was sufficient to allow digestion to proceed to completion. This reagent was included in all further digests in this study.

A sample of each crude PCR product was digested with \(Alul\) and \(HaeIII\). However, when compared to the fragments generated from the CL product, there were differences between those in the blood RNA and the CL samples. Upon close inspection it could be seen that some bands in the FRO and FRE \(Alul\) digest samples were of the same size as the main bands in the CL sample. Conversely, in the \(HaeIII\) digest, it appeared that a band in the CL sample mirrored a major product in the FRO and FRE samples. In each digest, these extra bands had a much reduced intensity compared to the bands that they mirrored. In each case, the FRO and FRE samples yielded identical patterns as expected (figure 10.2).

The conclusion from these digests was that the major product amplified from the CL sample was the correct one and that another major product had been amplified in the FRO and FRE samples which was of a size sufficiently similar to the correct fragment as not to be distinguished on an agarose gel. It was thought likely that since the CL sample was skin fibroblast derived, there was an abundance of type III collagen which masked a lower abundance target sequence. However, it was probable that in the RNA samples derived from blood, the abundance of the two transcripts was more equal or biased against the type III transcript. In this case, it could be supposed that the other transcript may be amplified as well as, or in preference to that expected, hence the unexpected digestion fragments in the FRO and FRE samples.

In order to investigate the identity of the predominant PCR species amplified from blood, a Southern hybridisation experiment was performed. In addition to the blood- and cell line-derived RNA samples, type III transcripts from lymphoblast RNA were amplified and blotted. Lymphoblast cells are abnormal cells that can be found in the blood and blood-forming organs in individuals with lymphoblastic leukaemia. Amplification of
Total RNA from fresh (FRE) and frozen (FRO) blood and a cultured fibroblast cell line (CL) was reverse transcribed and PCR amplified with type III collagen primers as outlined in the text. The *AlaI* and *HaeIII* digestion fragments generated from each amplification product are presented. Fragments are in base pairs. M represents marker fragments. Resulting fragments are denoted by ‘C’ if they are derived from the cell line amplification product and ‘F’ if they are derived from the blood (fresh and frozen) amplification product. Actual sizes are not given since their relevance is limited in this context.
transcripts from these cells was effected in an identical manner to that of the FRE, FRO and CL samples using the
nested priming approach. Details of all the samples blotted are given in table 10.2. The genomic amplified
fragments were included to act as controls for the type III sequence expected.

The sequences on the gel were hybridised to the SS-819 fragment isolated from the probe pH-75 (figure
10.1). The probe pH-75 DNA (Dalgleish, unpublished data) was derived from clones pH-42 (Kirk, Elliot and
Dalgleish, unpublished data) and pH-33 (Mishulin et al., 1986) and covers 4.4 kb of the type III collagen cDNA.
Following low stringency washes, both the COL3FL/COL3FR-amplified and reamplified CL samples gave a good
signal at the expected position with respect to the position of size markers (results not shown). The CL sample
reamplified with the nested primers (COL3A1-L and COL3A1-R) only gave a low intensity signal and appeared to
generate a number of fragments one of which appeared to be of the expected size i.e. 224 bp. The only other sample
to give a signal (other than the control sequences) was the lymphoblast sample reamplified with the nested primers.
However, this was of low intensity and the multiple bands were of unexpected sizes. None of the blood samples
yielded a signal suggesting that the major fragment that had been amplified was not type III collagen sequence.

In order to identify the nature of the product from lymphoblast RNA, both strands of this sample were directly
sequenced as in section 2.12.1 using end-labelled COL3A1-L and COL3A1-R in separate reactions. The sequence
of the CL-derived product was determined, in the same manner, simultaneously. The sequence derived from the CL
sample was that of the region of the type III collagen cDNA expected. However, that derived from the lymphoblast
was unidentifiable when compared to sequences in the EMBL primate sequence database. Results are not shown.

10.4 CONCLUSIONS AND FURTHER WORK

Since the fourth step in the overall approach plan for this study was unsuccessful, the work could not
proceed further until the problems had been identified. It appears from the above, often conflicting, results that type
III collagen sequences were not amplified from RNA isolated from fresh and frozen blood or cultured lymphoblast
cells although the cell line RNA sample yielded correct sequences. This could have been due to an abundance of
cells, other than those possessing the desired collagen transcripts, in the blood sample which may have contained a
greater abundance of the mRNAs. This may have had the effect of masking the type III collagen mRNA species
from the desired cells. The solution to this would be to isolate each cell type from the total blood sample, and to
isolate RNA from each and analyse them individually. Following this, the suggested approach for the amplification
of sequences could be re-started and hopefully go to completion. This, however, does not explain the results
obtained when lymphoblast RNA was investigated. Other groups have been able to detect ectopic transcription in
such cells; Chan and Cole (1991) stated that lymphoblastoid cells produced ‘minute amounts’ of each of α1(I),
α2(I), α1(III) transcripts as well as those of α1(II) mRNA. They were able to identify an α1(II)Gly997Ser
substitution in a patient with spondyloepiphyseal dysplasia using both fibroblast and lymphoblastoid cells.
Unfortunately there was not sufficient time to conduct such an experiment but results from other laboratories are
encouraging in that collagen transcripts have been isolated and successfully amplified. One such example is that of
<table>
<thead>
<tr>
<th>SAMPLE NUMBER</th>
<th>SAMPLE IDENTIFICATION</th>
<th>substrate</th>
<th>primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>marker fragments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>reverse transcription negative control</td>
<td></td>
<td>COL3FL/COL3FR</td>
</tr>
<tr>
<td>2</td>
<td>PCR amplification negative control</td>
<td></td>
<td>COL3FL/COL3FR</td>
</tr>
<tr>
<td>3</td>
<td>RNA from fresh blood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>RNA from frozen blood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>RNA from lymphoblast cell line</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>RNA from fibroblast cell line</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>sample 1</td>
<td></td>
<td>COL3FL/COL3FR</td>
</tr>
<tr>
<td>8</td>
<td>sample 2</td>
<td></td>
<td>COL3FL/COL3FR</td>
</tr>
<tr>
<td>9</td>
<td>sample 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>sample 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>sample 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>sample 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>water</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>sample 2</td>
<td></td>
<td>COL3A1-L/COL3A1-R</td>
</tr>
<tr>
<td>15</td>
<td>sample 3</td>
<td></td>
<td>(nested reamplification)</td>
</tr>
<tr>
<td>16</td>
<td>sample 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>sample 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>sample 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>water</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>genomic from individual A</td>
<td></td>
<td>COL3A1-L/COL3A1-R</td>
</tr>
<tr>
<td>21</td>
<td>genomic from individual B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>genomic from individual C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>marker fragments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10.2  Details of type III collagen PCR amplified products analysed by Southern hybridisation.

PCR amplification samples which were subjected to Southern blot and hybridisation analysis are indicated. Three anonymous genomic DNA samples were also amplified to act as positive controls for hybridisation. Samples 13 and 19 were included to act as indicators of potential contamination of nested priming amplifications.
null alleles at COL1A1 being detected from peripheral blood lymphocytes of type I Of individuals (Virdi and Sykes, 1993).

An assumption that is made when considering the use of illegitimate transcripts from an alternative tissue source is that the mutant allele may not be expressed at the same basal rate as a normal allele. It would be necessary to investigate, therefore, different known mutations which result in a number of different molecular species to determine the limits of such an experimental approach to mutation detection.
CHAPTER ELEVEN

A STRATEGY FOR THE PRODUCTION OF A LIBRARY ENRICHED WITH HUMAN COLLAGEN SEQUENCES

11.1 INTRODUCTION

To date 19 collagen types have been characterised which are encoded by over 25 genes. Many of the collagen proteins were isolated prior to the discovery of the corresponding genes but more recently cDNA clones have been isolated before the protein; in fact proteins for some of the characterised cDNA sequences remain as yet undiscovered. The aim of this study was to generate a library enriched with collagen gene sequences which might lead to the characterisation of novel collagens.

There are two basic approaches to cloning specific genes: (i) shotgun cloning in which DNA (cDNA or genomic) is digested with a selected restriction enzyme, the fragments cloned non-selectively and recombinants screened with a probe related to the gene of interest, and (ii) specific cloning of a fragment by hybridising sequences of interest from the DNA digest with a purified probe for the gene. Obviously novel genes cannot be isolated by the second method but the former can be adapted for this purpose.

It is likely that if any novel collagens remain to be discovered, their expression is restricted to certain tissues in which they may appear as a low abundance protein either throughout the cell cycle or at specific developmental stages as for types II and IX, for example. This could prove to be a hindrance in the isolation of novel collagens. Genomic DNA provides two copies of each collagen gene and therefore each gene should have an equal probability of being isolated. However, if a full-length gene clone was desired, the corresponding cDNA would be a better substrate since collagen genes are large and of variable size and would invariably be fragmented and cloned in sections. To generate collagen cDNA, mRNA sequences must first be isolated and reverse transcribed. This process generates copies of all the genes in the mRNA population in proportion to the abundance of the mRNA. This does not generally present a problem in tissues known not to express such genes but in a cell that actively produces large amounts of one type of collagen (type I in skin, for example), the transcripts of these genes will be in greater abundance than those of lesser-transcribed genes and the latter will be masked.

A library produced from cultured skin fibroblast cDNA had already been produced in the laboratory (Molyneux, 1991). This had been made using a degenerate primer PCR approach employing oligonucleotides based on the repetitive sequence motif present in collagen coding sequences. However, of all the sequences that were subjected to colony hybridisation, only 2.7% yielded a positive signal after Southern blot analysis using a cocktail of fibrillar collagen cDNAs as probes. The sequences of these clones corresponded to regions of the probe DNAs. Since the method had proved to be successful, albeit inefficient, it was decided to attempt to improve upon the protocol in order to isolate novel collagen sequences. The polymerase chain reaction had been employed to generate the library using a degenerate priming system as opposed to conventional single locus PCR. There are a number of ways of producing such a library; the cloning of members of multigene families with degenerate PCR
primers has been discussed by Wilkie and Simon (1991) and Telenius et al. (1992). Lee et al. (1988) generated a mixed population of primers to allow for codon degeneracy to produce a cDNA probe which was used to isolate a clone containing a full-length porcine urate oxidase cDNA. As has been discussed in Chapter One, and in common with many multigene families, collagen genes exhibit regions of highly conserved amino acid sequence. In order to isolate novel members of such gene families, PCR DNA amplification can be biased towards the family by synthesising oligonucleotides that are degenerate for the repetitive region of the sequence. The resulting amplification products can be cloned, screened and selected recombinants sequenced. Telenius et al. (1992) termed this method degenerate oligonucleotide primer PCR (DOP-PCR).

In such an application, the products might represent novel collagenous sequences. However, they would not be full-length collagen clones and it would be necessary to utilise the sequence as a probe against cloned sequences from a conventionally produced library in a hybridisation experiment. This would also eliminate any sequence errors potentially introduced by Taq DNA polymerase in the PCR amplification process. This procedure would probably identify a number of contiguous collagenous sequences which could then be combined to determine the nature of the full-length collagen sequence. The novel, complete collagen sequence could then be investigated further.

11.2 DESIGN OF A DEGENERATE COLLAGEN PRIMER

The primer used in this project was based on that used in the previous study (Molyneux, 1991). This primer, RANcDNA-2, exploited the repetitive nature of the triple helical domain of human collagen genes and the frequency of occurrence of proline residues in the third position of the Gly-Xaa-Yaa repeat motif. Since proline was found to be the most commonly occurring amino acid in this position in the major fibrillar collagen genes, the primer had been designed to incorporate a proline codon and a number of degenerate positions. An EcoRI restriction enzyme site was incorporated into the 5' end of the primer to aid cloning of amplification products. In this study, in order to try and improve upon the content of the library, the primer was modified. The basic structure of RANcDNA-2 was thought to be suitable for the experiment but since the amplifications using the primer had yielded few collagen sequences, it was decided to alter it such that it might potentially anneal to a more specific template. The structure of RANcDNA-2 is illustrated in figure 11.1a.

11.2.1 The 3' end of the primer

The 3' end of the primer was modified in an attempt to increase the specificity of the annealing. A codon for a glycine residue must be resident in the first position thus initially, the suitability of the primer was re-examined in order to verify that proline was the correct amino acid codon to include at the third position. Previously (Molyneux, 1991), three collagen gene sequences — COL1A1, COL1A2 and COL3A1 — had been aligned using the LINEUP program in the GCG molecular biology package. The consensus sequence indicated that proline was indeed the most commonly occurring residue in the third position. In this study, α1(I), α2(I), α1(II), α1(III), α2(V), α1(VII) and α2(VIII) sequences were analysed. The analysis was extended to the X position amino
Figure 11.1: Degenerate oligonucleotides — structure and comparison.

The structure of the primers used in DOP-PCR is illustrated: (a) original degenerate primer, (b) modified primer, (c) non-complementary 'tag' sequence. Sequence in lower case characters represents sequence non-complementary to collagen sequence included to aid cloning of amplification products. N indicates that any base, A, C, G or T is equally likely to occupy the position. Restriction enzyme recognition sites are indicated. The amino acids encoded by the primer sequence are indicated by single letters in accordance with the standard code, N indicates that any amino acid is likely to occupy the space.
acids also. Since some of the sequences analysed were shorter than others, the number of times that each amino acid occurred in both the second and third positions was recorded and these figures were converted to a percentage of the total number of X or Y position amino acids recorded. The results are summarised in table 11.1. As can be seen, proline is the most abundant amino acid in both the Xaa and the Yaa positions in all sequences studied. Thus the consensus sequence region of the oligonucleotide as determined by Molyneux (1991), i.e. 5' GGNN NNN CCN GG 3', was not altered. However the one modification made was the doubling in length of this consensus sequence. It was thought that this might make the binding of the primer more specific to collagen sequences. The 3' end of the primer is the most crucial part for ensuring correct binding of the primer to complementary sequence (Sommer and Tautz, 1989) but it was still hoped that the extra 11bp would aid in the specificity of binding. In general, amplification via the PCR requires two oligonucleotides complementary to opposite strands of target DNA; however, the collagen consensus sequence is present in both DNA strands and thus it is not necessary to design both a coding and a non-coding primer. The single primer would anneal on either the sense or antisense DNA strand. This primer was compared to a subset of entries in the EMBL database comprising all known human collagen sequences listed in the database at the time. As expected there were many matches showing that the primer had the potential to anneal to collagen sequences.

11.2.2 The 5' end of the primer

The 5' end of the original primer had been designed with future cloning purposes in mind. An EcoRI restriction enzyme site had been included immediately 5' to the degenerate part of the primer and an additional 8bp of random sequence were incorporated into the 5' end of the enzyme site since it had been reported by Jung et al. (1990) that certain enzymes fail to digest when the recognition sequence is at the terminus of a stretch of DNA and Jeffreys et al. (1990) showed that EcoRI digested when an 8bp extension was incorporated. The new primer was altered such that, in addition to an EcoRI site, BamHI and SalI restriction sites were also included in this region. The full length of the non-complementary region was 21bp. The EMBL primate database was screened for this non-complementary sequence. Any matches found displayed low complementarity and were thus considered to be negligible. Thus the primer would not anneal to any other sequences in preference to collagen sequences; this was important since the rest of the primer was degenerate. It was hoped that this ‘tag’ sequence would increase the specificity of priming by the oligonucleotide such that a greater number of collagen sequences could be amplified. It was proposed that a second primer be made with the same sequence as the tag portion of the degenerate primer. In the initial amplification steps the degenerate primer would bind to complementary sequences on the target DNA; in subsequent steps, the tag would be used to amplify these sequences. This, in theory, should prevent the oligonucleotides binding to repeated internal sites in the amplification products and would increase the abundance of the originally amplified sequences. Thus two primers were designed, DEG-cDNA-1 and DEG-TAG, DEG-TAG being of identical sequence to the non-complementary region of the degenerate primer. The structures of these two primers are illustrated in figure 11.1b and 11.1c and can be compared with the structure of RANcDNA-2. Since there are 10 degenerate positions in the DEG-cDNA-1 primer, there are a possible $4^{10}$ different oligonucleotides in
<table>
<thead>
<tr>
<th>PROTEIN SEQUENCE ANALYSED</th>
<th>THREE MOST ABUNDANT AMINO ACIDS IN X POSITION</th>
<th>NUMBER OF OCCURRENCES</th>
<th>NUMBER AS A PERCENTAGE OF ALL X POSITION AMINO ACIDS</th>
<th>THREE MOST ABUNDANT AMINO ACIDS IN Y POSITION</th>
<th>NUMBER OF OCCURRENCES</th>
<th>NUMBER AS A PERCENTAGE OF ALL Y POSITION AMINO ACIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>α1(I)</td>
<td>PROLINE</td>
<td>118</td>
<td>35.1</td>
<td>PROLINE</td>
<td>115</td>
<td>51.3</td>
</tr>
<tr>
<td></td>
<td>ALANINE</td>
<td>60</td>
<td>17.9</td>
<td>ALANINE</td>
<td>58</td>
<td>25.9</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>42</td>
<td>12.5</td>
<td>ARGinine</td>
<td>41</td>
<td>18.3</td>
</tr>
<tr>
<td>α2(I)</td>
<td>PROLINE</td>
<td>109</td>
<td>32.4</td>
<td>PROLINE</td>
<td>93</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>ALANINE</td>
<td>48</td>
<td>14.3</td>
<td>ALANINE</td>
<td>60</td>
<td>17.5</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>42</td>
<td>12.5</td>
<td>ARGinine</td>
<td>41</td>
<td>12.0</td>
</tr>
<tr>
<td>α1(II)</td>
<td>PROLINE</td>
<td>79</td>
<td>33.2</td>
<td>PROLINE</td>
<td>78</td>
<td>32.6</td>
</tr>
<tr>
<td></td>
<td>ALANINE</td>
<td>41</td>
<td>17.2</td>
<td>ALANINE</td>
<td>30</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>35</td>
<td>14.7</td>
<td>ARGinine</td>
<td>30</td>
<td>12.6</td>
</tr>
<tr>
<td>α1(III)</td>
<td>PROLINE</td>
<td>100</td>
<td>28.1</td>
<td>PROLINE</td>
<td>152</td>
<td>43.2</td>
</tr>
<tr>
<td></td>
<td>ALANINE</td>
<td>48</td>
<td>13.4</td>
<td>ALANINE</td>
<td>42</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>46</td>
<td>12.9</td>
<td>ARGinine</td>
<td>39</td>
<td>11.1</td>
</tr>
<tr>
<td>α2(V)</td>
<td>PROLINE</td>
<td>86</td>
<td>31.3</td>
<td>PROLINE</td>
<td>98</td>
<td>35.9</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>31</td>
<td>11.3</td>
<td>ALANINE</td>
<td>35</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>ASPARTIC ACID</td>
<td>20</td>
<td>7.3</td>
<td>ALANINE</td>
<td>21</td>
<td>7.7</td>
</tr>
<tr>
<td>α1(VII)</td>
<td>PROLINE</td>
<td>16</td>
<td>28.6</td>
<td>PROLINE</td>
<td>28</td>
<td>50.0</td>
</tr>
<tr>
<td></td>
<td>LEUCINE</td>
<td>8</td>
<td>14.3</td>
<td>ALANINE</td>
<td>9</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>7</td>
<td>12.5</td>
<td>ARGinine</td>
<td>4</td>
<td>7.1</td>
</tr>
<tr>
<td>α2(VIII)</td>
<td>PROLINE</td>
<td>44</td>
<td>29.9</td>
<td>PROLINE</td>
<td>74</td>
<td>52.5</td>
</tr>
<tr>
<td></td>
<td>LEUCINE</td>
<td>27</td>
<td>18.4</td>
<td>ALANINE</td>
<td>14</td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td>GLUTAMIC ACID</td>
<td>15</td>
<td>10.2</td>
<td>ALANINE</td>
<td>12</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Table 11.1  Amino acid frequencies in Xαα and Yαα positions of the collagen triple helix

The frequency of the three most commonly occurring amino acids in the Xαα and Yαα positions in the triple helix are given for seven collagen chains. The figures are indicated both as actual figures and as percentages of total Xαα or Yαα position residues.
the primer population. The theoretical mode of amplification by DEG-cDNA-1 and DEG-TAG is shown in figure 11.2.

11.3 THE AMPLIFICATION PROCEDURE

In the previous production of a library (Molyneux, 1991) standard PCR amplification of the substrate had been carried out i.e. repeated denaturation, annealing and extension steps at the same temperature each time. An annealing temperature of 55°C had been used in order to allow the degenerate primer to bind with an extension time of 10 minutes. In a further attempt to increase the specificity of the amplification reaction touchdown PCR was used. This method, described by Don et al. (1991), is useful for primers for which a dissociation temperature is incalculable. Since primer DEG-cDNA-1 used here has a high degree of degeneracy, the melting temperature could not be calculated with any degree of accuracy.

The template chosen for the initial experiment was genomic DNA isolated from the blood of a normal individual. Genomic DNA was chosen initially because of the equal number of copies of each gene present. It was hoped that this would act as a control for the amplification system and not be selective for genes of common collagen proteins.

Various ratios of DEG-cDNA-1 to DEG-TAG were investigated in order to determine the optimum levels of DEG-cDNA-1 with respect to DEG-TAG for amplification. The concentration of DEG-TAG was kept constant in all amplification reactions i.e. 1 μM final concentration. The DEG-cDNA-1 primer was used at a final concentration below that of DEG-TAG primer. The concentration of DEG-cDNA-1 was not increased above that of DEG-TAG since it was thought that under such conditions the degenerate primer might out-compete the tag primer in the latter rounds of amplification. A number of controls were also included; details of the reactions are illustrated in table 11.2. The DNA was amplified, using the touchdown approach, initially for two cycles consisting of 95°C for 1 min, 60°C for 1 min and 70°C for 10 min. The extension step was allowed to proceed for 10 minutes; it had been suggested by Innis et al. (1988) that the amplification of 1kb of sequence takes approximately one minute although Rychlik et al. (1990) suggested that the amount of sequence generated in this time is less. The amplification of 10kb in one reaction is unlikely, but to ensure that the extension time was not a limiting factor in the generation of long collagen sequences, the time was set at 10 minutes. Subsequent rounds of amplification consisted of cycles in which the annealing temperature was lowered by one degree every two cycles; the denaturation and extension conditions remained constant. The lowering of the annealing temperature was stopped when it reached 50°C, at which temperature 10 cycles of amplification were carried out. This resulted in a total of 30 cycles of amplification. One fifth volume of each of the amplification products was electrophoresed through a 0.8% agarose gel. This gel was Southern blotted and hybridised with a cocktail of three collagen cDNA probes namely α1(I), α2(I) and α1(III) sequences. Under low stringency wash conditions, a smeared signal in the lane corresponding to sample 7 (table 11.2) was observed (these results are not shown since the signal was particularly faint even after a long exposure to X-ray film; a photograph would not have been easy to reproduce). This was not unexpected since it was hoped that a population of sequences heterogeneous in length would be obtained from the
Figure 11.2  Mode of amplification of collagen sequences with the degenerate oligonucleotide DFG-cDNA-1

The theoretical mode of amplification of triple helical collagen sequences using degenerate oligonucleotides is illustrated. The primer is represented by the upper sequence in each case and is shown annealing to both the antisense and sense DNA strands.
Table 11.2 Conditions for genomic touchdown DOP-PCR.

The conditions used for the PCR are indicated. The cycle conditions are indicated in the text. Sample 10 represents the negative DNA control.

<table>
<thead>
<tr>
<th>REACTION NUMBER</th>
<th>AMOUNT OF DNA ng</th>
<th>FINAL DEG-TAG CONCENTRATION μM</th>
<th>FINAL DEG-cDNA-1 CONCENTRATION</th>
<th>DEG-TAG:DEG-cDNA-1 RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>300</td>
<td>-</td>
<td>1 μM</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>300</td>
<td>-</td>
<td>100 nM</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>-</td>
<td>10 nM</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>300</td>
<td>-</td>
<td>1 nM</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>300</td>
<td>-</td>
<td>1 μM</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>300</td>
<td>1</td>
<td>1 μM</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>300</td>
<td>1</td>
<td>100 nM</td>
<td>1:10</td>
</tr>
<tr>
<td>8</td>
<td>300</td>
<td>1</td>
<td>10 nM</td>
<td>1:100</td>
</tr>
<tr>
<td>9</td>
<td>300</td>
<td>1</td>
<td>1 nM</td>
<td>1:1000</td>
</tr>
<tr>
<td>10</td>
<td>-</td>
<td>1</td>
<td>1 μM</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 11.3 Size-selected amplification products.

Details of the samples removed from a low gelling temperature gel following electrophoresis of the degenerate amplification of genomic DNA.

<table>
<thead>
<tr>
<th>SAMPLE NUMBER</th>
<th>APPROXIMATE SIZE RANGE OF SELECTED FRAGMENTS bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;281</td>
</tr>
<tr>
<td>2</td>
<td>281-603</td>
</tr>
<tr>
<td>3</td>
<td>603-1078</td>
</tr>
<tr>
<td>4</td>
<td>1078-1353</td>
</tr>
<tr>
<td>5</td>
<td>1353-2027</td>
</tr>
<tr>
<td>6</td>
<td>2027-2322</td>
</tr>
<tr>
<td>7</td>
<td>&gt;2322</td>
</tr>
</tbody>
</table>
amplification rather than a small number of identical sequences. It appeared that the most effective ratio of 
DEG-cDNA-1 to DEG-TAG was 1:10 as judged by degree of hybridisation. The experiment was repeated using 
genomic DNA as a template and this optimal primer ratio. One fifth volume of the product was electrophoresed 
through a 0.8% agarose gel; this was to be cloned. Since smaller products will ligate more readily than larger ones 
it was decided to size select the product smear. This serves to maximise the number of different sized inserts within 
the population of clones. The smear was size selected using the 4X174RF/HaeIII and λHindIII digest markers as 
reference points on an agarose gel. Seven size ranges were isolated and these are described in table 11.3. The DNA 
in all isolated samples was purified as described in section 2.2.2 and resuspended in a final volume of 40 µl dH2O.

A fraction of each purified DNA sample was electrophoresed through an agarose gel alongside a fraction of the 
original product from which the size selected samples originated. This was in order to check that the isolation had 
been efficient and that the DNA was the expected size in each sample; both of these points were confirmed (results 
not shown). One half of each of the purified DNA samples was digested with EcoRI and ligated to EcoRI-digested, 
phosphatased pTZ19R plasmid vector; suitable controls were included. Transformation of the ligation sample, 
potentially containing sequences between 603bp and 1078bp (sample 3; table 11.3), into E. coli JM83 (recA) was 
carried out using the method of electroporation (Dower et al., 1988) since this is more efficient than a chemical 
transformation and hence would also increase the probability of detecting collagen sequences in the library. Only 
one sample was transformed initially to see if the method had been successful, the expectation was that at least one 
collagen sequence would be found.

11.4 RESULTS

The transformation yielded a large number of white colonies (indicative of plasmids potentially 
harbouring an insert) which were picked into microtitre plates as described in section 2.9. Colony blots were 
prepares and the DNA was hybridised to a cocktail of collagen probes which included cDNA sequences from 
a(I), α2(I), α1(III), α2(V), α1(VI) and α2(VI) collagen genes. A low stringency wash was carried out and the 
filters were autoradiographed at -80°C for two hours. A record was made of the 73 colonies that had produced a 
signal; these were then subgrouped into (i) colonies which had given a strong signal — 32 colonies— and 
(ii) colonies which had given a moderate signal — 41 colonies — with respect to the background signal 
determined from an area of one of the filters which harboured no colonies. The filters were then subjected to 
stringent washes. It was expected that some of the signals would become weaker with respect to the other colonies 
which would signify that these were non-recombinants. Those that retained a strong signal with respect to the other 
colonies might represent recombinants. However, following autoradiography, it was observed that the majority of 
the signals remained at the same relative intensities. It was decided to investigate some of these colonies to 
determine the nature of the insert. The 32 colonies which gave the strongest signals were used to inoculate 
overnight cultures from which the plasmid was isolated, as were 5 that gave a signal equivalent to that of the 
background to act as controls. Each of the plasmids was digested with EcoRI and the products analysed on an 
agarose gel. The results looked promising since 19 of the plasmids appeared to contain an insert of a size within the
expected range. The colonies which gave a background signal yielded no insert as expected. Two clones possessed an insert of slightly less than 600bp but this could be due to some smaller fragments being trapped with larger ones during electrophoresis of the PCR product prior to its size selection, imprecise removal of the agarose from the gel, collapse of the insert subsequent to cloning or simply the removal of 14bp of linker at each end during the procedure. Fresh digests were carried out; the samples were electrophoresed through a 0.8% agarose gel and subjected to Southern blot hybridisation analysis. The DNA was hybridised to a probe containing α1(I) and α2(I) cDNA sequences. A band representing the insert sequence from one of the clones termed G7:AE4 hybridised to the cDNA probe. This clone was sequenced using the universal M13 reverse sequencing primer as in section 2.12.2. Approximately 150bp of sequence were obtained from the clone (figure 11.3). Unfortunately, the sequence was not of a collagen or a collagen-like gene. Although it possessed a high degree of repetitive sequence units, it was more indicative of a minisatellite sequence than collagen. A search of the EMBL primate sequence database indicated that the best homology occurred between the newly identified sequence and involucrin. It was decided that this potential minisatellite could be investigated further.

11.5 INVESTIGATION OF THE CLONE G7:AE4

As can be seen from figure 11.3, there are noticeably repetitive sequence units. The limits of each 17 base pair block are indicated and it can be seen that each of the blocks is very similar with only minor differences. It is possible that the primer annealed to the sequence within repeat unit I since there is very good similarity between the sequence of the clone and the primer at the 3' end. However, the primer may well have annealed to bases further upstream of the G7:AE4 sequence which were not cloned. The amino acid sequence in all three frames and in both directions of the sequence of G7:AE4 was determined. It could be seen that although short stretches of Gly-Xaa-Yaa repeats were present they were no more extensive than three or six residues, in the main. Thus it was thought that the sequence was unlikely to be a collagen gene. However, due to the presence of repeated sequence motifs, minisatellite characteristics of the cloned sequence were investigated. Minisatellites have to meet certain criteria before they are considered suitable for use in diagnostic applications such as fingerprinting or MVR-PCR. Two of these criteria are internal sequence variation and a high allelic heterozygosity. From the sequence data, it can be seen that the DNA possesses a degree of sequence difference. To investigate the heterozygosity of the potential minisatellite, its occurrence in a large family of individuals was determined. Genomic DNA from a 17-member family obtained from the CEPH, Paris was used for the study. Each sample was digested with Adal, an enzyme with frequently occurring restriction sites and that does not cut within the G7:AE4 sequence. The samples were electrophoresed through a 1.0% agarose gel which was then Southern blotted. The G7:AE4 sequence was isolated from the cloning vector by digestion with EcoRI; this isolated fragment was used as a probe in the subsequent hybridisation analysis. Following a high stringency post-hybridisation wash step, a number of fragments were visualised, however, each member of the family analysed possessed fragments of the same size (results not shown). From this result it was concluded that the sequence would not be an informative minisatellite in diagnostic studies.
Figure 11.3  Sequence of the G7:AE4 clone

DNA sequence of the sense strand of the clone G7:AE4 is presented. The repetitive nature of the sequence is evident. The arbitrary start point of each repeat unit is indicated by an upwardly pointing arrow on the sequence ladder. Sequence prior to unit 1 is plasmid derived. Each repeat motif is numbered and the corresponding base sequence is written to the right. Repeat unit number 1 is the most 3' unit with respect to the start of the sequence. Bases which are underlined in the written sequence represent minor sequence differences between repeat units. The consistent difference of bases 10 and 11 of each unit (CA or GG) is not indicated.
I  GGACAGCCAGGCCACCT
II  GGAGCACCACACCACCC
III  GGAGCACCAGGCCACCC
IV  GGAGCACCACACCACCC
V  GGAGCACCAGGCCACCC
VI  GGAGCACCACACCACCC
VII  GGAGCACCACACCACCC
VIII  GGAGCACCCTGCCACCC
IX  GGAGCACCAGGCCACCC
X  AGAGCACCAGGCCACCT
11.6 DISCUSSION AND FUTURE WORK

Despite the adaptations made to the existing degenerate oligonucleotide and the techniques used to try and increase the specificity of amplification and the efficiency of cloning, no collagen sequences were isolated from this library. This was unexpected. It was hoped that some collagen would be present even if it was already characterised since the previous application of the technique had yielded fibrillar collagen sequences (Molyneux, 1991). There are several possible explanations for the failure of this approach. In this application, genomic DNA was used as a template as opposed to human fibroblast cDNA as in the previous library. Since the reasoning behind this was to decrease the possible 'masking' effect of the more abundant collagen cDNAs, it is possible that it was too effective and target sequences were at too low a concentration to be detected using this degenerate system. A useful experiment might have been the amplification of a known collagen cDNA/genomic sequence with the degenerate primer, DEG-cDNA-1, to see if a product was obtained. This experiment would have given an idea of whether the strategy as a whole would have been successful. Only products from one of the seven ligations were cloned (number 3, table 11.3). The length of insert in this sample was expected to be between approximately 603bp and 1078bp. It was thought unlikely that collagenous sequences of this length would not exist. This was illustrated by the hybridisation of touchdown PCR products with fibrillar collagen probes. Thus the transformation from this ligation should not have reduced the chances of finding collagenous sequences. Problems could have arisen due to the nature of the amplification technique. The primer was designed to anneal to the Gly-Xaa-Yaa motifs in collagenous domains. As has been discussed (Chapter One), other proteins have collagenous domains but possess none of the other properties of collagens and are not, therefore, classed as such. Such sequences could have been amplified by the PCR. However, these would also have been detected at the screening step as one colony losing intensity with respect to others following a high stringency wash. This effect was not seen at all and the removal of some colonies for analysis may have been sufficiently restricted such that, unknowingly, the non-collagen sequences were overlooked.

A second problem that may have occurred during amplification is one in which the oligonucleotide annealed to sequences flanking an intron. In this instance, the amplification product would consist predominantly of intronic sequence which, using the screening procedure of this project, i.e. the use of a cocktail of cDNA probes, would not hybridise. In collagen genes, introns are of large and variable size and the majority remain uncharacterised. Since so few have been sequenced it is not known if any share identity between collagen genes and therefore the substitution of a genomic collagen probe for the cDNA ones would probably not increase the likelihood of detecting a similar sequence. Yet another possibility is that the primer could have annealed to the terminal end of the collagenous domain (COL) of a FACIT. If the non-collagenous domain (NC) was subsequently amplified, the cDNA probes would not detect the sequence. In each of these scenarios, the assumption is that the primer would have annealed to collagenous sequences adjacent to the opposite end of the intron or NC domain. However, introns are large and in such a situation the length of DNA to be copied may exceed the limits of the PCR in many instances. NC domains are of variable size also. In the a(CXVI) gene most of the domains are short (11–39 residues) the exception being the large (312 residues) NC11 domain (Pan et al., 1992). Thus such an amplification could potentially occur in a number of instances in FACITs.
In addition, if the primer annealed at the end of the triple helical domain and DNA was synthesised towards the N-terminus, the lack of a priming site in the opposite direction would result in failure of amplification. Since time was a limiting factor in this project, further investigations were limited. However, one preliminary experiment carried out was the amplification of fibroblast and lymphoblast mRNA. Since the previous application had used fibroblast mRNA as a template, it was decided to see how the new degenerate oligonucleotide priming system worked on this substrate as well as lymphoblast mRNA. A number of reactions were set up and these are illustrated in Table 11.4. Following the PCR amplification, samples were analysed on an agarose gel. It could be seen that all the negative control samples were indeed devoid of product whereas it appeared that all the other reactions had worked to varying degrees. Samples, excluding the negative controls were electrophoresed through a fresh agarose gel (Figure 11.4a) and were Southern blotted. Hybridisation analysis consisted of annealing with a cocktail of fibrillar collagen probes (see above). Following exposure it could be seen that three samples had yielded a signal, namely F, G and K (Figure 11.4b). From these limited data it cannot be concluded that any amplification system is necessarily better than another since each of the three samples that gave a signal had been produced by different approaches. The best signal was obtained from the amplification of fibroblast mRNA with the primers used in touchdown PCR but with the alternative amplification conditions. The signal obtained from the two lymphoblast mRNA amplifications cannot be directly compared to that of the fibroblast sample since, as was seen in Chapter Ten, the abundance of different transcripts in these cells is unknown. However, the two lymphoblast samples are interesting in that samples F and K only differ in the nature of the final amplification step. The signals are obviously very different and it appears that touchdown PCR may be selecting for the amplification of larger fragments.

However, this was only a preliminary experiment and insufficient time meant that it was not repeated. It would be interesting to repeat the experiment to see if the same results were obtained. If the signals were reproducible it would be possible to clone the products of amplification in the same manner as described earlier in this chapter. It would be particularly interesting to see if cloned collagen sequences could be obtained from the lymphoblast reactions. If such sequences could be attained from the mRNA of these cells, they might be useful in the basal transcription study after all (Chapter Ten).
Table 11.4 Details of degenerate oligonucleotide amplification of fibroblast and lymphoblast RNA

Details of mRNA substrates and the mode of subsequent amplification are given for the preliminary follow-up experiment for production of a cDNA library of collagen sequences. FB, cultured fibroblast cell line; LB, cultured lymphoblast cell line. Where 'RT blank' or 'PCR blank' is given, the substrate for the reaction was water to act as a negative control. The primers given in the table are those discussed in the text. The PCR amplification approach is termed “touchdown” if the conditions outlined in the text for touchdown PCR were used and “55°C” if the conditions used were 35 cycles consisting of 95°C for 1.5 min, 55°C for 1.5 min and 70°C for 10 min.

<table>
<thead>
<tr>
<th>SAMPLE IDENTIFICATION</th>
<th>mRNA SUBSTRATE</th>
<th>PRIMERS USED IN REVERSE TRANSCRIPTION</th>
<th>PRIMERS USED IN PCR AMPLIFICATION</th>
<th>AMPLIFICATION APPROACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>FB</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>touchdown</td>
</tr>
<tr>
<td>B</td>
<td>FB</td>
<td>DEG-cDNA-2</td>
<td>RANcDNA-2</td>
<td>touchdown</td>
</tr>
<tr>
<td>C</td>
<td>(RT blank)</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>touchdown</td>
</tr>
<tr>
<td>D</td>
<td>(PCR blank)</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>touchdown</td>
</tr>
<tr>
<td>E</td>
<td>LB</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>touchdown</td>
</tr>
<tr>
<td>F</td>
<td>LB</td>
<td>RANcDNA-2</td>
<td>RANcDNA-2</td>
<td>touchdown</td>
</tr>
<tr>
<td>G</td>
<td>FB</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>55°C</td>
</tr>
<tr>
<td>H</td>
<td>FB</td>
<td>RANcDNA-2</td>
<td>RANcDNA-2</td>
<td>55°C</td>
</tr>
<tr>
<td>I</td>
<td>FB</td>
<td>DEG-cDNA-1</td>
<td>DEG-TAG</td>
<td>55°C</td>
</tr>
<tr>
<td>J</td>
<td>LB</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>55°C</td>
</tr>
<tr>
<td>K</td>
<td>LB</td>
<td>RANcDNA-2</td>
<td>RANcDNA-2</td>
<td>55°C</td>
</tr>
<tr>
<td>L</td>
<td>LB</td>
<td>DEG-cDNA-1</td>
<td>DEG-TAG</td>
<td>55°C</td>
</tr>
<tr>
<td>M</td>
<td>(RT blank)</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>55°C</td>
</tr>
<tr>
<td>N</td>
<td>(PCR blank)</td>
<td>DEG-cDNA-1</td>
<td>DEG-cDNA-1 + DEG-TAG</td>
<td>55°C</td>
</tr>
</tbody>
</table>
Figure 11.4  Preliminary investigation of fibroblast and lymphoblast RNA as a source of collagen transcripts for the production of a collagen cDNA library

(a) PCR amplification products corresponding to those given in table 11.4 are presented. Control samples were not included (see text).

(b) Results of the Southern blot and hybridisation of the gel in (a) are presented. Signals can be seen in lanes G and K and to a lesser extent in lane F.

Fragments are in base pairs.
CHAPTER TWELVE

CONCLUDING REMARKS

One of the aims of this study was to identify and characterise mutations in COL1A2 that could result in the OI phenotypes of the patients under investigation. The approach taken was SSCP analysis coupled with DNA sequence determination. Four novel mutations were identified each of which altered a glycine residue of the triple helical domain to another, bulkier amino acid. These have been summarised and discussed in Chapter Nine. Three of these mutations altered a glycine to a serine and added to the limited number of such substitutions reported for the α2(I) chain of type I collagen. The data also indicated that they do not conform to a simple gradient model of a genotype to phenotype relationship and so reflect the similar situation observed in the α1(I) chain for these types of substitution. Two of the glycine by serine substitutions were identified in more than one unrelated individual. There have only been a limited number of so-called shared OI mutations identified to date and these novel substitutions added to those previously reported. Each of these mutations was a G to A transition occurring at a CpG dinucleotide supporting data that suggest that sequences such as these represent mutation hot-spots. In addition, the glycine by serine substitution found in a single individual is the most N-terminal of any substituted residue identified in the α2(I) chain to date. These substitutions gave rise to a spectrum of OI phenotypes. The fourth mutation characterised is novel in that it represents the first glycine by glutamic acid substitution found in either of the type I collagen genes which is responsible for an OI phenotype. The lethal phenotype of the patient was not unexpected when the amino acid change was likened to reported glycine by aspartic acid substitutions. In some cases, protein data were provided by collaborating laboratories which localised the mutations to a specific region of the triple helical domain. In such cases, the point of mutation as determined by SSCP and DNA sequencing analyses was compared. In one instance out of four, the protein and DNA data did not agree suggesting that cyanogen bromide peptide mapping data are not fully informative.

In addition to the mutations that cause OI phenotypes, two single base changes that did not alter the amino acid sequence were identified. One altered a BglII restriction enzyme site. However, they each only occurred in single individuals and as such were deemed uninformative for linkage studies. A previously characterised PvuII RFLP was also detected.

Not all of the patients studied were found to harbour a sequence defect responsible for the OI phenotype using this approach. The triple helix of both type I collagen α-chains had been screened in this and in a previous study (Mackay, 1992) and hence an estimation of the efficiency of the method with respect to the detection of OI mutations could be made. Since this figure was only approximately 60%, different approaches to mutation detection were investigated utilising knowledge of OI mutations gained during SSCP analysis. The more extensive study involved looking for altered restriction enzyme sites that could be indicative of OI mutations, however, no novel mutations were identified although a positive control was seen to result in an altered band pattern. Having identified potential problems associated with this approach, a modification was made to the system which involved
the two-dimensional separation of the digested DNA. However, there was insufficient time in which to develop this fully. It is hoped that the system could be developed in the laboratory and that further mutations in the population of patients studied here could be identified. If successful it is thought that this method could be applied to both type I collagen genes in the search for OI mutations and could possibly be adapted for the detection of mutations in other genes e.g. COL3A1 in EDS IV patients. Because of the many variables that are important for determining phenotype, rapid methods for the identification and characterisation of mutations would be of benefit in providing clinical information to families and would assist in understanding the manner in which mutations resulted in the clinical manifestations of OI.

Unfortunately, at present, there is no cure for OI and thus the identification of mutations underlying the phenotype may appear to be of purely academic interest. However, the identification of the mutation in a family with recurrent OI could enable screening for parental mosaicism and may be of help to the parents when considering future pregnancies. An important use of such accumulated mutation data could be in the identification of families with osteoporosis which, if detected at an early stage, may be correctable.

Another aim of the study was to identify a suitable source of basal level collagen gene transcripts for use in mutation analysis which would also circumvent the need for invasive methods of tissue sampling. Despite initial experiments yielding promising results, the source chosen for study, namely blood, proved to be unsuitable with the approach taken. However, it is hoped that with modifications to the strategy, such transcripts could be amplified. It had been hoped that if this project had been successful, blood mRNA could also be used as a source for cDNA from which a collagen library could be generated. The initial approach taken for the production of a collagen-enriched library was to redesign a degenerate PCR oligonucleotide with which collagenous triple helical regions could be amplified. The ability of the primer to amplify collagen sequences was investigated using genomic DNA as a template. From the resulting amplification product, sequences were cloned into a plasmid vector and a potential positive sequenced. The sequence illustrated that it was GC-rich but was not collagen-like in its identity. It could be speculated where the primer had annealed and thus show that the actual amplification process had been successful albeit not specific for collagen. It would be interesting to see if modifications to the basal transcription study could yield a suitable template for use in the production of the library.

The isolation of novel collagens may be important in the identification of genes underlying a number of known heritable connective tissue disorders.
REFERENCES


procollagen to aspartate in a lethal variant of osteogenesis imperfecta.


J. Biol. Chem. 263: 7855–7861


Nucl. Acids Res. 19:4193–4198

Chelly, J., Hugnot, J.-P., Concordet, J.-P., Kaplan, J.-C., Kahn, A. (1991a) Illegitimate (or ectopic) transcription proceeds through the usual promoters.

J. Clin. Invest. 88: 1161–1166


Nature 333: 858–860

J. Biol. Chem. 268: 18218–18225

J. Biol. Chem. 266: 12487–12494


BioTechniques 15: 586


*XLIIIth Meeting of the Federation of European Connective Tissue Societies, Davos (Switzerland), July 12–17, 1992.* Abstract 21/3


*Annu. Rev. Biochem.* 53: 717–748


*Hum. Mut.* 2: 380–388


*Proc. Natl. Acad. Sci. USA* 78: 7360–7364


*Hum. Genet.* 88: 508–512


*Nucl. Acids Res.* 18: 3933–3939


*Proc. Natl. Acad. Sci. USA* 90: 10325–10329


*J. Biol. Chem.* 266: 12035–1204


*J. Biol. Chem.* 264: 9632–9637


*Hum. Genet.* 91: 175–177


*Anal. Chem.* 62: 1202–1214


*Fifth International Conference on Osteogenesis Imperfecta, Oxford (UK) September 27–30, 1993 Abstract:* 120


*Nucl. Acids Res.* 18: 4227–4236


*Genomics* 12: 836–837
Grobler-Rabie, A.F., Wallis, G., Breyten, D.K.,
Detection of a high frequency Rsal polymorphism in
the human proo2(I) collagen gene which is linked to
an autosomal dominant form of osteogenesis
imperfecta.
EMBO J. 4: 1745-1748

mutations in nucleic acids.
Nature Genet. 5: 111-117

Guenet, J.L., Stanesco, R., Maroteaux, P., Stanesco,
recessive mutation in the mouse.
J. Hered. 73: 440-441

single-stranded DNA by the polymerase chain
reaction and its application to direct sequencing of
the HLA-DQA locus.

Handt, O., Reis, A., Schmidtke, J. (1992) Ectopic
transcription of the parathyroid hormone gene in
lymphocytes, lymphoblastoid cells and tumour
tissue.
Journal of Endocrinology 135: 249-256

Harbers, K., Kuehn, M., Delius, H., Jaenisch, R.
(1984) Insertion of retrovirus into the first intron of
α1(I) collagen gene leads to embryonic lethal
mutation in mice.
Proc. Natl. Acad. Sci. USA 81: 1504-1508

3: Evolution of the extracellular matrix in
invertebrates.
FASEB J. 7: 1115-1123

Retrovirus insertion inactivates mouse α1(I)
collagen gene by blocking initiation of transcription.
Nature 320: 365-367

Hawkins, J.R., Superti-Furga, A., Steinmann, B.,
COL1A1 in a lethal variant of osteogenesis
imperfecta.
J. Biol. Chem. 266: 22370-22374

Hayashi, K. (1991) PCR-SSCP: A simple and
sensitive method for detection of mutations in the
genomic DNA.
PCR Methods and Applications 1: 34-38

PCR-SSCP?
Hum. Mut. 2: 338-346

Helmingen, H.J., Kiraly, K., Peltturi, A., Tammi,
M.I., Vandenbergh, P., Pereira, R., Dhillipala, R.,
Khillan, J.S., Ala-Kokko, L., Hum, E.L., Sokolov,
B.P., Prockop, D.J. (1993) An inbred line of
transgenic mice expressing an internally deleted
gene for type II procollagen (COL2A1). Young
mice have a variable phenotype of a
chondrodysplasia and older mice have osteoarthritic
changes in joints.
J. Clin. Invest. 92: 582-595


*J. Biol. Chem.* 262: 5540–5545


*Am. J. Hum. Genet.* 47: 112–120


*Nucl. Acid Res.* 19: 3507–3510


*Cell* 57: 807–816

Structure and Function of Collagen Types.
Academic Press, Inc., Orlando pp. 1-42

Mutations in collagen genes: causes of rare and some common diseases in humans.  
FASEB J. 5: 2052–2060

A 19-base pair deletion in the pro-α2(I) gene of type I procollagen that causes in-frame RNA splicing from exon 10 to exon 12 in a proband with atypical osteogenesis imperfecta and in his asymptomatic mother.  
J. Biol. Chem. 263: 11407–11413

Structure of a full-length cDNA clone for the preproc2(I) chain of human type I collagen.  

A cysteine for glycine substitution at position 1017 in an α1(I) chain of type I collagens in a patient with mild dominantly inherited osteogenesis imperfecta.  
Mol. Biol. Med. 5: 197–207

Characterization of point mutations in the collagen COL1A1 and COL1A2 genes causing lethal perinatal osteogenesis imperfecta.  
J. Biol. Chem. 264: 15809–15812

Identification of the molecular defect in a family with spondylepiphysial dysplasia.  
Science 244: 978–980

Generation of cDNA probes directed by amino acid sequence: Closing of urate oxidase.  
Science 239: 1288–1291

J. Med. Genet. 30: 951–954

A procedure for DNA and RNA transfer to membrane filters avoiding weight-induced gel flattening.  

Type I procollagen containing a substitution of serine for glycine 883 in the proα1(I) chain, in an individual with mild osteogenesis imperfecta, is cleaved normally by N-proteinase and does not have a kink at the site of the substitution.  
Int. J. Exp. Path. 74: A11–A12

Type I procollagens containing substitutions of aspartate, arginine, and cysteine, for glycine in the proα1(I) chain are cleaved slowly by N-proteinase, but only
the cysteine substitution introduces a kink in the molecule.

*J. Biol. Chem.* 267: 25521–25528


Mackay, K. (1992) "Molecular analysis of type I collagen genes in inherited disorders.” Ph.D., University of Leicester


*Hum. Mut.* (in press)


*Hum. Mol. Genet.* 2: 1155–1160

Mackay, K., Lund, A.M., Raghunath, M., Steinmann, B., Dalgleish, R. (1993b) SSCP detection of a Gly565Val substitution in the proα1(I) collagen chain resulting in osteogenesis imperfecta type II.

*Hum. Genet.* 91: 439–444


*Genomics* 13: 809–813


*British Journal of Ophthalmology* 77: 515–524


McKusick, V.A., (1956) Heritable Disorders of Connective Tissue. C.V. Mosby, St. Louis


Molyneux, K., Sturman, B.J., Byers, P.H., Dalgleish, R. (1993) A single amino acid deletion in
the α2(I) chain of type I collagen produces osteogenesis imperfecta type III.

*Hum. Genet.* 90: 621–628


Fifth International Conference on Osteogenesis Imperfecta, Oxford (UK) September 27–30, 1993 Abstract: 129


*Arch. Biochem. Biophys.* 182: 33–41


*Hum. Mut.* 2: 196–204


*J. Biol. Chem.* 258: 10128–10135


*Nature* 313: 495–498

Myers, R.M., Fischer, S.G., Lerman, L.S., Maniatis, T. (1985b) Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis.

*Nucl. Acids Res.* 13: 3131–3145


Niyibizi, C., Bonadio, J., Byers, P.H., Eyre, D.R. (1992) Incorporation of type I collagen molecules that contain a mutant α2(1) chain (Gly580→Asp) into bone matrix in a lethal case of osteogenesis imperfecta. J. Biol. Chem. 267: 23108-23112


Olsen, A.S., Geddes, A.E., Prockop, D.J. (1990) High levels of expression of a minigene version of the human proα1(I) collagen gene in stably transfected mouse fibroblasts. J. Biol. Chem. 266: 1117-1121


Phillips, C.L., Shrago-Howe, A.W., Pinnell, S.R., Wensnap, R.J. (1990) A substitution at a non-glycine position in the triple-helical domain of...
pro2(I) collagen chains present in an individual with a variant of the Marfan syndrome.


J. Biol. Chem. 259: 12941–12944


Poschl, E., Pollner, R., Kühn, K. (1988) The genes for the α1(IV) and α2(IV) chains of human basement membrane collagen type IV are arranged head-to-head and separated by a bidirectional promoter of unique structure.

EMBO J. 7: 2687–2695


Hum. Mut. 2: 192–195


Adv. Prot. Chem. 35: 1–104

Prockop, D.J. (1990) Mutations that alter the primary structure of collagen.

J. Biol. Chem. 265: 15349–15352


New Engl. J. Med. 34: 12–23


Hum. Genet. 87: 33–40

Pruchno, C.J., Willing, M.C., Starman, B.J., Aylesworth, A., Byers, P.H. (1989) Inefficient splicing and production of both a normal and shortened mRNA from the same COL1A1 allele of type I collagen in a father and son with osteogenesis imperfecta (OD).

Am. J. Hum. Genet. 46: 213 (abstract 837)


FASEB J. 4: 1616–1623


Rychlik, W., Rhoads, R.E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. 

*Nucl. Acids Res.* 17: 8543-8551


*Nucl. Acids Res.* 18: 6409-6412


*Am. J. Hum. Genet.* 49: 797-803


*Science* 239: 487-491


*Science* 230: 1350-1354


*Hum. Mut.* 1: 63-69


*Proc. Natl. Acad. Sci. USA* 74: 5463-5467

Sarker, G., Sommer, S.S. (1989) Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. 

*Science* 244: 331-334


*Nucl. Acids Res.* 20: 871-878


*Genomics* 13: 441-443


*Nucl. Acids Res.* 19: 674


D-periodic distribution of collagen type IX along cartilage fibrils.

*J. Cell. Biol.* **106**: 991-997


*Science* **237**: 415-417


*Gene* **19**: 2989-2998


*Fifth International Conference on Osteogenesis Imperfecta, Oxford (UK) September 27-30, 1993* Abstract: 148


*J. Biol. Chem.* **264**: 18265-18267

Vogel, B.E., Doetz, R., Kadler, K.E., Hojima, Y., Engel, J., Prockop, D.J. (1988) A substitution of cysteine for glycine 748 of the α1 chain produces a kink at this site in the procollagen I molecule and an altered N-proteinase cleavage site over 225 nm away.

*J. Biol. Chem.* **263**: 19249-19255

Vogel, B.E., Minor, R.R., Freund, M., Prockop, D.J. (1987) A point mutation in a type I procollagen gene converts glycine 748 of the α1 chain to cysteine and destabilizes the triple helix in a lethal variant of osteogenesis imperfecta.

*J. Biol. Chem.* **262**: 14737-14744


*Annu. Rev. Biochem.* **59**: 837-72


*J. Med. Genet.* **30**: 494-496


*J. Biol. Chem.* **267**: 25529-25534

Wallis, G.A., Starman, B.J., Chessor, S.D., Willing, M.C., Byers, P.H. (1990a) Mutations in the CB6 and carboxyl-terminal regions of COL1A1 causing lethal OI have different effects on the stability and secretion of the type I collagen molecule.

*Fourth International Conference on Osteogenesis Imperfecta, Pavia (Italy) September 9-12, 1990.* Abstract: 55


Westerven, R., Shamaj, J., Prockop, D.J. (1990) Mutations that substitute serine for glycine \( \alpha_1(1)-598 \) and glycine \( \alpha_1(1)-631 \) in type I procollagen. *J. Biol. Chem.* 265: 13995–14000


de Wet, W., Chu, M.-L., Prockop, D.J. (1983a) The mRNAs for the pro-\( \alpha_1(1) \) and pro-\( \alpha_2(1) \) chains of type I procollagen are translated at the same rate in normal human fibroblasts and in the fibroblasts from two variants of osteogenesis imperfecta with altered steady state ratios of the two mRNAs. *J. Biol. Chem.* 258: 14385–14389.

de Wet, W., Pihlatiemi, T., Myers, J., Kelly, T.E., Prockop, D.J. (1983b) Synthesis of a shortened pro-\( \alpha_2(1) \) chain and decreased synthesis of pro-\( \alpha_2(1) \) chains in a proband with osteogenesis imperfecta. *J. Biol. Chem.* 258: 7721–7728


Youssoufi, H., Kazazian Jr, H.H., Phillips, D.G.,
Aronis, S., Tsiftis, G., Brown, V.A., Antonarakis,
S.E. (1986) Recurrent mutations in haemophilia A
give evidence for CpG mutation hotspots.
Nature 324: 380-382

Zafarullah, K., Kleinert, C., Tronp, G., Kuivaniemi,
H., Koutumaari, S., Wu, Y., Ganguly, A., Prockop,
D.J. (1990) G to A polymorphism in exon 31 of the
COL3A1 gene.
Nucl. Acids Res. 18: 6180

Zheng, H., Hasty, P., Brenneman, M.A., Grompe,
Fidelity of targeted recombination in human
fibroblasts and murine embryonic stem cells.

Zhou, J., Smeets, B., Mochizuki, T., Antignoc, C.,
Identification of a novel type IV collagen gene, α-6(IV),
and characterisation of deletions in patients
with X-linked Alport syndrome and
leiomyomatosis.

Zhou, J., Hostikka, S.L., Chow, I.T., Tryggvason,
K. (1991a) Characterisation of the 3' half of the
human type IV collagen α5 gene that is affected in
the Alport syndrome.
Genomics 9: 1-9

Zhuang, J., Constantinou, C.D., Ganguly, A.,
Prockop, D.J. (1991) A single base mutation in type
I procollagen (COL1A1) that converts glycine alpha
1-541 to aspartate in a lethal variant of osteogenesis
imperfecta: detection of the mutation with a
carbodiimide reaction of DNA heteroduplexes and
direct sequencing of products of the PCR.
Am. J. Genet. 48: 1186-1191

Zhou, J., Barker, D.F., Hostikka, S.L., Gregory,
M.C., Atkin, C.L., Tryggvason, K. (1991b) Single
base mutation in α5(IV) collagen chain gene
converting a conserved cysteine to serine in Alport
syndrome.
Genomics 9: 10-18

Zhou, J., Smeets, B., Mochizuki, T., Antignoc, C.,
Identification of a novel type IV collagen gene, α-
6(IV), and characterisation of deletions in patients
with X-linked Alport syndrome and
leiomyomatosis.

Zhou, J., Hostikka, S.L., Chow, I.T., Tryggvason,
K. (1991a) Characterisation of the 3' half of the
human type IV collagen α5 gene that is affected in
the Alport syndrome.
Genomics 9: 1-9

Zhuang, J., Constantinou, C.D., Ganguly, A.,
Prockop, D.J. (1991) A single base mutation in type
I procollagen (COL1A1) that converts glycine alpha
1-541 to aspartate in a lethal variant of osteogenesis
imperfecta: detection of the mutation with a
carbodiimide reaction of DNA heteroduplexes and
direct sequencing of products of the PCR.
Am. J. Genet. 48: 1186-1191
A novel glycine to glutamic acid substitution at position 343 in the α2 chain of type I collagen in an individual with lethal osteogenesis imperfecta

Nicola J. Rose, Katrina Mackay, Peter H. Byers* and Raymond Dalgleish*

Department of Genetics, University of Leicester, Leicester LE1 7RH, UK and Departments of Pathology and Medicine, University of Washington, Seattle, WA 98195, USA

Received July 21, 1993. Revised and Accepted October 26, 1993

© 1993 Oxford University Press

Human Molecular Genetics, 1993, Vol. 2, No. 12 2175—2177

The lethal form of osteogenesis imperfecta (OI type II) has been estimated to affect between 1 in 20,000 to 1 in 60,000 individuals (1), and is characterised by dark sclerae, a beaked nose, a very soft calvarium and short limbs with bowed legs (2). The ribs are beaded and the thoracic cavity is often very small, this making respiratory insufficiency a common cause of death. Congestive heart failure and infection are other causes of mortality and individuals with type II OI rarely survive beyond one month after birth, with approximately 60% of cases being fatal in the first 24 hours (1).

In virtually all instances, OI type II results from a dominantly inherited mutation of one of the type I collagen genes (COL1A1 or COL1A2) that is altered to encode a bulkier amino acid. This new amino acid interferes with the winding of the three proα-chains into the procollagen molecule (3). Point mutations in the first position of the GGN glycine codon would result in codons for arginine, serine, cysteine, tryptophan or a stop codon, whereas those in the second position would result in codons for alanine, aspartic acid or glutamic acid. To date, the substitution of glycine by glutamic acid has not been observed in reported cases of OI. Here we describe the detection of a heterozygous G to A transition in COL1A2 which caused the glycine at position 343 of the triple helix of α2(I) to be substituted by glutamic acid and resulted in the lethal variant of OI in the proband. The convention of designating the first amino acid of the triple helix as number one has been followed throughout.

The proband (our cell strain 89-208) was the first child born to phenotypically normal non-consanguinous parents (cell strains 89-247 and 89-248). The child had characteristic features of the perinatal lethal form of OI (soft calvarium, small thorax, short limbs) and the radiological picture included virtual absence of calvarial mineralization, beaded ribs, platyspondyly, short accordan-like femurs, marked tibial bowing, and generalized hypomineralization of the skeleton. The infant died shortly after delivery. Dermal fibroblasts were grown from skin punch biopsies obtained from the affected infant and from both parents, using standard conditions (4). Cells from the affected infant synthesized and secreted a diminished amount of normal type I procollagen molecules although the difference is not marked. The cells retained a species of type I procollagen in which both the prox(I) and prox(2I) chains had delayed electrophoretic mobilities (Figure 1A). This is confirmed following partial pepsin digestion of the molecules (Figure 1B). Relative to the levels of α1(III) trimer, there is a clear reduction of α1(I) and α2(I) chains in the medium of the patient’s cells as compared with the control. The delayed mobility of the abnormal peptides with respect to the control was evident and very few of the abnormal molecules appeared to be secreted. The cyanogen bromide peptides obtained from the amino-terminal end of the triple helix of the abnormal α1(I) chains within the cells also had delayed electrophoretic mobilities (Figure 1C). In the medium it is apparent that the α1(I)CB8 peptide is tilted upwards and merges with the normal α1(I)CB7 fragment. The α1(I)CB8-3 is also tilted slightly upwards, an effect which is more apparent in the cell layer with respect to fragments α1(I)CB7-6 and CB3-7. This is compatible with overmodification in the domain amino-terminal to residue 401 of the triple helix, suggesting that the mutation lay between residues 122 and 401 of the triple helix (exons 13—25) of either type I collagen gene. Parental fibroblasts synthesized only normal type I procollagen molecules.

The mutation was localized by single strand conformation polymorphism (SSCP) analysis (5, 6), using amplified cDNA fragments of the COL1A1 and COL1A2 genes of type I collagen. RNA was isolated from confluent fibroblasts by the method of Chomczynski and Sacchi (7) and synthetic oligonucleotide primers were used to amplify, by PCR, overlapping fragments of the entire cDNA sequence of both genes for SSCP analysis. No abnormalities were detected in the entire COL1A1 cDNA sequence by this method. In contrast, the fragment amplified by primers A2A (8) (sequence 5' TCCCTGGACCCCGGTCTGTTTCTGAG - 3') and SSCP1A2-Z (sequence 5' GGGTTACCAATITCACCTCTG - 3') contained electrophoretic variants. Following first strand cDNA synthesis (9), PCR amplification was carried out as described by Kogan et al. (10) for thirty-five cycles consisting of 95°C for 1.2 min, 55°C for 1.0 min and 70°C for 1.2 min. Four double restriction enzyme digests (HinfI/Styl, HindIII/Didel, HindIII/MspI and Alul/Didel) were performed on the amplification product and SSCP analysis was performed essentially as described by Mackay et al. (11). Fragments were separated by electrophoresis on a 0.5% Hydrasil-MDE gel (AT Biochem, Malvern, PA) containing 10% glycerol and 0.6×TBE (54mM Tris, 54mM boric acid, 1.2 mM EDTA; pH 8.3) at room temperature and at 3W. Fragments with altered mobilities were observed in the...
Direct sequence determination of a PCR-generated cDNA template (14) using a collagen sequence specific internal oligonucleotide [SEQA2-21; 5'-CTCCAGGACCTCCTGGGC-.

Hinfl/Styl, Hinfl/ Ddel and Hinfl/Mspl digests corresponding to the 189 bp non-denatured Hinfl fragment in each instance (Figure 2). A corresponding fragment with altered mobility was not observed in the fourth digest (Alul/Ddel) possibly due to the nature of the 'conformer' produced by the fragment. We conclude that, in the context of the restriction enzyme digest with Alul and Ddel, the mutation does not produce a conformer whose mobility is altered with respect to that of the wild type gene. It has been noted previously that not all fragments known to or assumed to harbour a sequence change have an altered mobility with respect to that of the wild type gene. It has been noted previously that not all fragments known to or assumed to harbour a sequence change have an altered mobility with respect to that of the wild type gene.

Figure 1. Procollagens, collagens and cyanogen bromide peptide maps of collagens. A. Proteins in the medium and in the cell layer were separated by SDS–PAGE under reducing condition. The cells from the patient (O1) retain type I procollagen, the chains of which have abnormal electrophoretic mobilities. FN; fibronectin. B. Following partial proteolytic digestion with pepsin the α1(I) and α2(I) chains from the abnormal molecules are seen to be selectively retained within the cells. C. The peptides from (B) were cleaved in the gel with cyanogen bromide and separated by electrophoresis in the second dimension. The medium samples are on the left and the cell-layer samples on the right. Only the mobility of the α1(III)CB8 peptide from the abnormal chain is shifted upward (most apparent in the cell-layer sample as the partial product CB8-3) indicating that the overmodification produced by the mutation begins in that domain, between amino acids 122 and 401 of the triple helix. The bar is extended arbitrarily through α1(IIICB8) because it is difficult to estimate from the extent of the modification where in the peptide the overmodification begins.

Figure 2. SSCP analysis of the patient. Non-denatured samples (N) are identified by upper case characters, 'A' being the largest fragment from each digest. Corresponding denatured samples (D) are identified by lower case characters. The band shifts caused by the base change are denoted by 'c' in both the Hinfl/Ddel and Hinfl/Styl digests and by 'b' in the Hinfl/Mspl digest. In each case, lane 1 represents the patient sample and lanes 2, 3 and 4 represent control samples. Not all fragments are shown.

Figure 3. DNA sequencing of the sense strand of double stranded PCR products showing the region surrounding the mutation in patient R and a control individual (C). The mutant sequence displays a G to A transition. An AvaII restriction enzyme site destroyed by the mutation is indicated by an asterisk and bar.

The previous failure to identify substitution of a glycine residue by glutamic acid (encoded by GAA or GAG) in the chains of
type I collagen possibly reflects two factors: (i) the paucity of the mRNA of the triple helical domain and (ii) the potentially of the Avai sites within the intron were not determined.

lines represent the exons (E). Known fragment sizes are indicated, the positions derived from the intronic sequence. B. An Avai restriction map of the genomic parents and control. The three bands whose size can only be approximated are 91 bp fragments indicative of the mutation is present in the proband but not the are shown for the father (F), proband (P), mother (M), and a control (C). The

of the triple helical domain. The latter possibility was investigated to a glutamic acid codon might only be expected in 3.2% of «1(1)... As no previous data exist, the phenotypic consequences of 45 bp — — 46 bp

to a glutamic acid codon might only be expected in 3.2% of «1(1) and 883 (19). In the «2(1) chain, similar substitutions occur (which has a shorter side chain), it can be postulated that glycine by glutamic acid substitutions might result in a lethal phenotype wherever they occur along the triple helical domain.

ACKNOWLEDGEMENTS

Drs K. Molyneux and J.R. Hawkins are acknowledged for their help with tissue culture and the SSCP technique respectively. We thank Drs Charles Winston and Helga Toriello for providing clinical information about and cells from the proband studied here. This work was supported by a grant from the Medical Research Council, who also supported NJR with a studentship, and by grants from the National Institutes of Health (AR 21557 and AR 41223).

REFERENCES