RECOMBINANT HUMAN COLLAGENS

Characterization of type II collagen expressed in insect cells and production of types I-III collagen in the yeast Pichia pastoris

MINNA NOKELAINEN

Department of Medical Biochemistry, University of Oulu
MINNA NOKELAINEN

RECOMBINANT HUMAN COLLAGENS
Characterization of type II collagen expressed in insect cells and production of types I-III collagen in the yeast Pichia pastoris

Academic Dissertation to be presented with the assent of the Faculty of Medicine, University of Oulu, for public discussion in the Auditorium of the Department of Medical Biochemistry, on September 14th, 2000, at 1 p.m.

OULUN YLIOPISTO, OULU 2000
Nokelainen, Minna, Recombinant human collagens. Characterization of type II collagen expressed in insect cells and production of types I-III collagen in the yeast Pichia pastoris

Collagen Research Unit, Biocenter Oulu and Department of Medical Biochemistry, University of Oulu, FIN-90220 Oulu, Finland

2000

Oulu, Finland

(Manuscript received 10 August 2000)

Abstract

An efficient system for expressing recombinant human collagens is expected to have numerous scientific and medical applications, but this is difficult to achieve because most systems do not have sufficient levels of activity of prolyl 4-hydroxylase, the key enzyme of collagen synthesis.

A recombinant form of human type II collagen, the main structural component of cartilage, was produced here in insect cells by coinfecting them with two baculoviruses, one coding for the proα chains of human type II procollagen, and the other for both the α and β subunits of human prolyl 4-hydroxylase. The amino acid composition of the recombinant form was very similar to that of the non-recombinant protein, with the exception that the hydroxylysine content was very low. The highest expression levels obtained in suspension cultures were 50 mg/l. An additional baculovirus coding for human lysyl hydroxylase was used to express type II collagen with a high hydroxylysine content. Marked differences in the rate of fibril formation in vitro and the morphology of the resulting fibrils were found between the recombinant type II collagens having 2 and 19 hydroxylysine residues/1000 amino acids, the maximal turbidity of the former being reached within 5 min, whereas the absorbance of the latter increased up to about 10 h. In addition, the latter collagen formed thin fibrils, whereas the former produced thick fibrils on a background of thin ones. The data indicate that regulation of the extent of lysine hydroxylation, and consequently of the amounts of hydroxylysine-linked carbohydrate units, may have major effects on collagen fibril formation.

In order to study the expression of recombinant human collagens in yeasts, cDNAs for the proα chains of procollagens of type I, II and III were transformed into a recombinant P. pastoris strain expressing human prolyl 4-hydroxylase subunits. All the P. pastoris strains obtained produced full-length proα chains. Cells coexpressing the proα1(I) chains and prolyl 4-hydroxylase produced homotrimeric type I procollagen molecules, whereas cells coexpressing the proα1(I) and proα2(I) chains and prolyl 4-hydroxylase produced heterotrimeric molecules with the correct 2:1 chain ratio. pCα1(I) and pCα2(I) chains lacking the N propeptides assembled into pCcollagen molecules and yielded correctly folded and fully hydroxylated collagen molecules upon pepsinization. The Tm values of recombinant type I-III collagens produced in shaker flasks were about 38°C and the degree of hydroxylation of proline residues was lower than that in the corresponding non-recombinant collagens. When the recombinant collagens were produced in a 2-litre fermentor equipped with an O2 supply system, the expression levels increased markedly to 0.2 – 0.6 g/l. In addition, all these collagens were identical in 4-hydroxyproline content to the corresponding non-recombinant proteins, and all of them formed native-type fibrils.

Keywords: procollagen, prolyl 4-hydroxylase, recombinant protein expression
Acknowledgements

The work was carried out at the Collagen Research Unit of the Department of Medical Biochemistry, University of Oulu, and Biocenter Oulu.

I wish to express my deepest gratitude to my supervisor, Professor of the Academy of Finland Kari Kivirikko, for giving me the opportunity to prepare my thesis as a member of his research group. His knowledge of collagens and their modifying enzymes and his never-ending optimism and support have been of great value to me and have been essential for the completion of this work. I also owe my sincere gratitude to my additional supervisor, Docent Johanna Myllyharju, for her valuable contribution and guidance at all stages of the work. I also wish to express my appreciation to Professor Taina Pihlajaniemi, for her help with all manner of things and for providing excellent research facilities. My sincere thanks are also due to Professor Ilmo Hassinen and Docent Leena Ala-Kokko.

I wish to thank Docent Volkmar Güntzler and Professor Aarno Palotie for their valuable comments of the manuscript of this thesis and also Malcolm Hicks, M.A., for his careful revision the language of the manuscript and of the original articles.

I am also very grateful to all my co-authors of the articles, especially Tarja Helaakoski, Holger Notbohm, Honming Tu and Annamari Vuorela, for their invaluable contributions to this work.

My friends and colleagues at the Department of Medical Biochemistry deserve my warmest thanks. The fruitful and relaxing scientific and non-scientific conversations with my roommates Pirkko Huhtala, Ritva Nissi, Kaisa Passoja and Kati Rautavuoma have been of great value to me. I also wish to thank Kaisa Passoja, Anne Snelman and Tero Pihlajamäki for their friendship and support in completing this thesis and in all practical arrangements when organizing the dissertation party.

I wish to extend my sincere thanks to all the staff of the Department of Medical Biochemistry. Especially I wish to acknowledge the skillful technical help and assistance of Raija Juntunen, Anne Kokko and Tanja Väisänen and their positive attitude towards this work. I am grateful to Marja-Leena Kivelä and Auli Kinnunen for their skillful secretarial assistance and for their kind help in many daily problems. The Computer – Team, Ari-Pekka Kvist, Juha Näräkkangas and Marko Kervinen, are acknowledged for
their invaluable help. Pertti Vuokila has been of great help when ordering reagents, as has been Seppo Lähdesmäki with the fermentor and many other technical problems.

I am deeply grateful to all my friends outside the scientific world for giving me a chance to think and talk about something else than the collagens. My parents, my sister and my mother-in-law deserve my special thanks for their care and their encouraging attitude towards my studies.

Finally, I owe my deepest gratitude to my husband, Pasi, for his love, understanding and support during these years. I am grateful to him for his scientific knowledge and above all for sharing an interesting life outside research work.

This work was supported financially by FibroGen Inc, the Research and Science Foundation of Farmos and the Finnish Centre of Excellence Programme (2000-2005) of the Academy of Finland (44843).

Oulu, August 2000

[Signature]
Abbreviations

\( \alpha \) chain          a collagen polypeptide chain
BMGY                      buffered complex-glycerol medium
BMMY                      buffered minimal methanol medium
bp                        base pair(s)
BSA                       bovine serum albumin
CD                        circular dichroism
cDNA                      complementary DNA
C propeptide              C-terminal collagen propeptide
ECM                       extracellular matrix
ER                        endoplasmic reticulum
kb                        kilobases
kDa                       kilodalton
MOI                       multiplicity of infection
\( M_r \)                  relative molecular mass
N propeptide              N-terminal collagen propeptide
PAGE                      polyacrylamide gel electrophoresis
PBS                       phosphate-buffered saline
pC\( \alpha \) chain        a procollagen polypeptide chain produced by cleavage of the N propeptide
PDI                       protein disulphide isomerase
4-PH                      prolyl 4-hydroxylase
pN\( \alpha \) chain        a procollagen polypeptide chain produced by cleavage of the C propeptide
\text{pro}\( \alpha \)(I) chain  pro\( \alpha \)1 chain of type I procollagen
\text{pro}\( \alpha \)(II) chain  pro\( \alpha \)1 chain of type II procollagen
\text{pro}\( \alpha \)(III) chain  pro\( \alpha \)1 chain of type III procollagen
\( T_m \)                  midpoint of thermal transition from helix to coil
X (in Gly-X-Y)             any amino acid
Y (in Gly-X-Y)             any amino acid
List of original articles

This thesis is based on the following articles, which are referred to in the text by their Roman numerals:


## Contents

Abstract
Acknowledgements
Abbreviations
List of original articles
Contents

1 Introduction ................................................................................................................... 13
2 Review of the literature ................................................................................................. 15
   2.1 The collagen family of proteins ................................................................. 15
      2.1.1 The structure and stability of the collagen triple helix .................... 16
      2.1.2 Fibril-forming collagens ........................................................................ 17
         2.1.2.1 Collagen types I, II and III ........................................................... 19
         2.1.2.2 Collagen types V and XI ............................................................... 21
   2.2 Biosynthesis of collagen .............................................................................. 22
      2.2.1 Hydroxylation of proline residues .................................................. 23
         2.2.1.1 Prolyl 4-hydroxylase ................................................................. 25
      2.2.2 Hydroxylation of lysine residues and glycosylation of hydroxylysine residues .......................................................... 26
         2.2.2.1 Lysyl hydroxylase ........................................................................ 27
      2.2.3 Procollagen assembly and secretion ................................................... 28
         2.2.3.1 Protein disulphide isomerase (PDI) ................................................ 29
      2.2.4 Procollagen processing and fibril assembly ...................................... 30
         2.2.4.1 Fibril formation ............................................................................. 31
   2.3 Expression of recombinant proteins in a baculovirus system ......................... 32
   2.4 Expression of recombinant proteins in the yeast *Pichia pastoris* ................. 34
   2.5 Recombinant human collagens in therapeutic applications ......................... 37
3 Outlines of the present research ............................................................................. 39
4 Materials and methods ........................................................................................... 41
   4.1 Expression of human type II collagen using a baculovirus expression system (I, II) .......................................................... 41
      4.1.1 Construction of baculovirus transfer vectors and generation of recombinant viruses (I, II) .......................................................... 41
4.1.2 Expression and analysis of recombinant proteins
in insect cell cultures (I, II)................................................................. 42
4.1.3 Purification and characterization of recombinant type II collagen (I, II) .... 42
4.1.4 Fibril formation experiments (II).................................................. 43
4.2 Expression of human collagen types I-III using a yeast expression system
(Pichia pastoris) (III, IV)............................................................................... 44
4.2.1 Construction of the P. pastoris expression plasmids and generation of
the recombinant P. pastoris strains (III, IV)................................................. 44
4.2.2 Expression and analysis of the recombinant prolyl 4-hydroxylase and
type I, II and III procollagens (III, IV)..................................................... 44
4.2.3 Fermentation of the recombinant P. pastoris strains (III, IV)................. 45
4.2.4 Characterization of recombinant proteins produced
in the 2-litre fermentor (III, IV)............................................................. 46
4.2.5 Purification and characterization of the recombinant collagens
produced in the 2-litre fermentor (III, IV)................................................... 46
5 Results............................................................................................................. 47
5.1 Expression of human type II procollagen in insect cells (I,II)....................... 47
5.1.1 Characterization of the recombinant human type II collagen (I, II)......... 48
5.1.2 Expression and characterization of recombinant human type II collagen
with increased hydroxylysine content (I, II)............................................. 49
5.1.3 Formation of type II collagen fibrils in vitro (II)...................................... 49
5.2 Expression of type I procollagen with and without its N propeptide in
Pichia pastoris in shaker flasks (III).......................................................... 50
5.3 Expression and characterization of type II and III procollagens
in Pichia pastoris in shaker flasks (IV)....................................................... 52
5.4 Production of recombinant human type I-III collagens in a 2-litre fermentor
and characterization of the recombinant proteins (III, IV)......................... 52
6 Discussion...................................................................................................... 54
6.1 Expression of type II collagen in insect cells and characterization of
the resulting protein................................................................................... 55
6.2 Expression of type I-III procollagens in Pichia pastoris and
characterization of the resulting proteins............................................... 57
7 References.................................................................................................... 60
1 Introduction

The extracellular matrix (ECM) is a structural and functional entity composed of a network of many macromolecules. Its components are insoluble fibrils, microfibrils, soluble proteins and glycoproteins. The ECM has an important role in providing tissues with their mechanical and physiological properties, and it also influences cell attachment and cell migration.

Collagen is the most abundant proteins of the ECM. Their function is to support the architecture of various tissues and organs. They are found in all tissues, but are particularly abundant in bone, tendon, skin, ligaments, cartilage and blood vessels. The most commonly occurring collagen types are types I, II and III, which form the long-recognized characteristic fibril bundles seen in many tissues. These bundles are formed by the self-assembly of triple helical collagen molecules, adjacent molecules being staggered with respect to each other and the fibrils being stabilized by covalent intermolecular cross-links. The self-assembly of collagen molecules into fibrils depends on their triple helical structure. The triple-helical domains of collagen have the repeating amino acid sequence (Gly-X-Y)$_n$, where X and Y are frequently the amino acids proline and hydroxyproline, respectively. Another characteristic feature of collagen is the post-translational enzymatic hydroxylation of specific proline and lysine residues and further modification by glycosylation of some of the hydroxylysine residues.

Collagen is now used commercially as biomaterials in numerous medical applications involving soft and hard tissue repair. They are also used as drug delivery systems, and all gelatines are prepared from collagens. Large amounts of collagens are currently required for these applications, the most important sources being bovine skin and bones. Since there is a risk that these bovine-derived collagens may carry disease-causing contaminants, an alternative source of collagens would be highly desirable. Recombinant human collagens would seem an excellent substitute for animal collagens in many applications, but their production presents major problems because of the post-translational modifications required to achieve fully folded, stable triple helical collagen molecules.

Prolyl 4-hydroxylase is one of the enzymes required for collagen biosynthesis, as it catalyses the formation of 4-hydroxyproline residues in the nascent procollagen chains. These 4-hydroxyproline residues are an essential requirement for the formation of a
stable triple-helical procollagen molecule as the procollagen chains do not form triple helices that are stable at body temperature in the absence of this enzyme activity. Lysyl hydroxylase catalyses the hydroxylation of lysine residues in procollagen chains, the resulting hydroxylysine residues acting as attachment sites for carbohydrate units and being essential for the formation of intermolecular cross-links in the collagen fibrils.

The recombinant baculovirus expression system in insect cells is one of the most commonly used systems for the expression of foreign genes in eukaryotic cells. This system is a helper-independent viral system which has been used to express heterologous proteins in large quantities from many sources. It has many advantages over other expression systems. Baculoviruses do not infect mammals or plants, and are thus safe to use. In addition, the level of expression of foreign proteins is usually high and the resulting recombinant protein contains appropriate post-translational modifications. Furthermore, the method is suitable for large scale production, as certain insect cell lines are easy to scale up. Several applications of baculovirus technology have now been developed and are commercially available.

The Pichia Expression System has become increasingly common for the production of high levels of functionally active recombinant proteins. *Pichia pastoris* is a methylotrophic yeast that can grow on methanol as its sole carbon source. As a eukaryotic organism, it is also capable of many of the post-translational modifications performed in higher eukaryotic cells. *P. pastoris* cultures can easily be scaled up to high cell densities, and in many cases the levels of expression of secreted or cytoplasmic heterologous proteins are higher than in any other recombinant protein production system.

The present study focuses on the expression of recombinant human type II collagen in insect cells and on the role of hydroxylysine residues in fibril formation. A further aim was to produce recombinant human type I-III collagens in the yeast *Pichia pastoris*, to scale up the production in a 2-litre fermentor, and to investigate the formation of fibrils composed of type I collagen homotrimers and heterotrimers *in vitro*. 
2 Review of the literature

2.1 The collagen family of proteins

The collagens form a family of proteins that constitute the major structural components of the ECM, representing approximately one-third of the body protein in man. They are found in almost all tissues of the body, and are particular abundant in bone, skin, tendon, ligaments, cartilage and vessel walls. At least 19 types exist, and there are more than 30 gene products (for reviews, see Kielty et al. 1993, Kivirikko 1993, Pihlajaniemi & Rehn 1995, Prockop & Kivirikko 1995, Bateman et al. 1996). The general definition of a collagen is that the protein consists of three identical or related polypeptide chains which are folded into at least one triple-helical domain and assemble into supramolecular aggregates in the extracellular space. Each polypeptide chain, called an α chain, is coiled into a helical conformation, and the three helical chains are wrapped around each other to form a triple helix. Triple helix formation requires the presence of glycine as every third residue, since this is the only amino acid that can fit into the restricted space available at the centre of the collagen helix. Another general feature of collagens is the presence of a high proportion of proline and lysine residues in the repeating -Gly-X-Y- sequences, with extensive co-translational and post-translational modifications involved in these residues (see Kielty et al. 1993, Kivirikko 1993, Prockop & Kivirikko 1995).

Collagens can be divided into two groups on the basis of their primary structure and macromolecular assembly, i.e. the fibril-forming and non-fibril-forming collagens. The fibril-forming collagens include types I, II, III, V and XI, which have a large, uninterrupted triple-helical domain capable of fibril formation. Their fibrils create the structural frameworks for many tissues. The non-fibril-forming collagens include types IV, VI-X and XII-XIX, and are a more heterogeneous group. One typical feature is the presence of one or more non-collagenous interruptions within the collagenous sequence. Members of this group form filaments, sheet-like structures, network-like elements and anchoring fibrils and can associate with collagen fibrils and membranes that are present in a highly tissue-specific manner in a variety organs. The collagen types are designated with Roman numerals in the order of their discovery, and the α chains found in each

Several other proteins exist that contain collagenous sequences but do not belong to the group of collagens. These proteins contain repetitive -Gly-X-Y- sequences and form triple-helical structures but are not structural components of the extracellular matrix and hence are not defined as collagens (see Pihlajaniemi & Rehn 1995, Prockop & Kivirikko 1995). The following sections will focus on the structures and biosynthesis of the fibril-forming collagens.

### 2.1.1 The structure and stability of the collagen triple helix

All collagen molecules consist of three polypeptide chains, each folded into a left-handed polyproline II helical conformation, and the three helical chains are then wrapped around each other into a right-handed triple helix. In some collagen types all three α chains in the molecule are identical (homotrimer), while in other types the molecule contains two or even three different α chains (heterotrimer) (see Kielty et al. 1993, Prockop & Kivirikko 1995).

The presence of glycine as every third amino acid in the repeating -Gly-X-Y- sequence in each of the chains is essential, because a larger amino acid would not fit into the restricted space available at the centre of the triple helix, where the three chains come together. Stabilization of the triple helix by glycines has been studied using collagen-like peptides (Bella et al. 1994, 1995), and a very large number of glycine substitutions in the collagen triple helix have been identified in various heritable connective tissue disorders (see Kivirikko 1993, Olsen 1995, Dalgleish 1998, De Paepe 1998 and references therein). The conformation of the triple helices formed by the collagen-like (Pro-4Hyp-Gly)10 peptides was altered when a Gly→Ala substitution was introduced at the centre of the peptide. This substitution resulted in a small local untwisting of the triple helix and reduced its thermal stability (Bella et al. 1994). Bella et al. (1994) concluded that similar conformational changes may occur in the case of the various Gly→X substitutions in collagens that lead to various disease manifestations. It has also been thought that glycine substitutions may reduce the rate of folding of the triple helix, as a Gly→Cys substitution within the collagen domain of type 1 collagen prolongs the time needed to reach the triple-helical state (Raghunath et al. 1994).

A triple-helical region has a high proline content in the X positions of the repeated Gly-X-Y sequence and a high 4-hydroxyproline content in the Y positions. These residues are required for the correct conformation of the helix, and the 4-hydroxyprolines are also essential for thermal stability of the helix (see Kielty et al. 1993, Prockop & Kivirikko 1995). The hydroxylation of proline residues increases the thermal stability of the triple helix and raises its melting temperature from about 23°C to 42°C (Berg & Prockop 1973, Rosenbloom et al. 1973). Steric constraints imposed by the imino acid rings of the proline and hydroxyproline residues favour triple helix formation because the five-membered rings of the imino acids are rigid and limit rotation of the N-C peptide bond (see Kielty et al. 1993, Prockop & Kivirikko 1995). Water molecules are also believed to have a
significant role in maintaining the conformation of collagen molecules (Doege & Fessler 1986). The crystal structure of a collagen-like peptide points to repetitive hydrogen-bonded interactions between the triple-helical peptides and water molecules (Bella et al. 1995, Kramer et al. 1999). 4-Hydroxyproline residues were shown to provide additional sites for hydrogen bonding with the peptide surface. This means that 4-hydroxyproline residues are critical for both the stability and the supramolecular structure of collagens (Bella et al. 1995, Kramer et al. 1998).

2.1.2 Fibril-forming collagens

The fibril-forming collagens, types I, II, III, V and XI, were originally grouped together as group I collagens, since they are similar in their protein structure. The main triple-helical domain forms a rod-like structure with a length of about 300 nm and a diameter of about 1.5 nm (see Mayne & Burgeson 1987, Miller & Gay 1987). These collagens form fibrils, which primarily act as supporting elements in the extracellular matrix. Their triple helical domains consist of about 1000 residues in uninterrupted -Gly-X-Y- sequences, but the molecules are first synthesized as larger precursors, termed procollagens, that contain large, non-collagenous N and C propeptide domains in addition to the triple-helical region. The N and C propeptides are linked to the main triple-helical domain by short, non-collagenous sequences, called telopeptides. The telopeptides are the primary sites for intermolecular cross-linking, which is important for the stabilization of the collagen fibres (Eyre et al. 1984). The C propeptide is highly conserved among the fibril-forming collagen types in terms of both its amino acid sequence and the predicted secondary structure, and contains about 250 amino acid residues (Dion & Myers 1987). The numbers and locations of the cysteine residues in the C propeptides are highly conserved. The four extreme C-terminal cysteines are invariant and form intrachain disulphide linkages, while the three or four remaining cysteines are involved in intermolecular disulphide bonds (Doege & Fessler 1986). In addition, the C propeptides of all the fibril-forming collagens except that of the proα2(XI) chain contain an Asn-X-Ser/Thr tripeptide acceptor sequence at amino acid position 153-155 for the attachment of an N-linked oligosaccharide unit (Kimura et al. 1989). These conserved features of the C propeptides are believed to reflect their essential roles in initiating chain assembly and triple helix formation. The N propeptide is more variable in structure and is between 50 and 500 residues in length (see van der Rest & Garrone 1991, Kiely et al. 1993). The proα1(I) and proα1(III) N propeptides are similar and comprise a globular cysteine-rich domain and a short triple helical domain (Tromp et al. 1988, Janeczko & Ramirez 1989). The proα2(I) N propeptide lacks the cysteine-rich domain (de Wet et al. 1987). Alternative splicing of the pre-mRNA for type II procollagen leads to two forms of the proα1(II) chain, one including the cysteine-rich globular domain and the other lacking in this (Ryan & Sandell 1990). The proα1(V), proα2(V) and proα1(XI) N propeptides are larger than those of the major fibril-forming collagens and their collagenous domains contain two interruptions in the Gly-X-Y sequence (Woodbury et al. 1989, Yoshioka & Ramirez 1990, Takahara et al. 1991). The procollagen molecules are processed to mature collagen...
molecules by cleavage of the N and C propeptides by specific N and C proteinases. Finally, a characteristic feature of the fibril-forming collagens is that they form highly ordered, quarter-staggered, 67 nm banded fibrils, i.e. adjacent molecules overlap by a distance of 67 nm or a multiple of this, with a 40 nm gap between the ends of the continuous non-overlapping molecules (for reviews, see Kiely et al. 1993, Prockop & Kivirikko 1995, Bateman et al. 1996).

![Diagram of collagen structure](image)

**Fig. 1. Structure of a type I procollagen molecule.** The white arrows indicate the sites for propeptide cleavage by specific N and C proteinases. Intrachain disulphide bridges (dotted lines) are present within both the N and C propeptides, whereas interchain disulphide bonds are present only between the C propeptides. One high-mannose, N-linked oligosaccharide unit is attached to each C propeptide, and some hydroxylysine residues in the collagen domain are further modified by the addition of galactose or galactose and glucose.

The genes encoding the fibril-forming collagens I-III show structural similarity but vary in size from 18 to 44 kb (Vuorio & de Crombrugghe 1990, Chu & Prockop 1993, Bateman et al. 1996). These genes consist of 51-53 exons, the major triple-helical domain being encoded by 44 exons. The similarities in gene structure also extend to the propeptides, the N propeptides being encoded by six exons in the genes for all forms, whereas the C propeptides are encoded by four exons. The gene loci for the members of the collagen family have been given names beginning with COL in a human context and col in animals, followed by an Arabic numeral denoting the collagen type, the letter A for an α chain, and another Arabic numeral for the α chain in question. For example, COL1A2 is the gene locus for the proα2(I) chain, and COL2A1 for the proα1(II) chain. All the exons encoding the triple-helical domain are of sizes that are multiples of 9 bp (e.g. 45, 54, 99, 108 and 162 bp, coding for 5, 6, 11, 12 and 18 Gly-X-Y triplets, respectively), the most common size being 54 bp. It has been suggested that the ancestral gene for the fibril-forming collagens must have evolved by amplification of a 54-bp unit embedded into intron sequences (Yamada et al. 1980). All these exons start with a codon for glycine and end with a complete codon for an amino acid in the Y position. Cloning...
of the COL5A1 and COL11A2 genes has indicated, however, that they did not evolve with the genes for the major fibril-forming collagens, since marked differences are found in exon sizes, codon usage and the number of exons coding for the triple-helical region (Takahara et al. 1995, Vuoristo et al. 1996).

Table 1. Fibril-forming collagens in human tissues.

<table>
<thead>
<tr>
<th>Type</th>
<th>Chain</th>
<th>Molecular forms</th>
<th>Mainly expressed</th>
<th>Key Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>α1(I)</td>
<td>[α1(I)]2α2(I)</td>
<td>Widely distributed; dermis, bone, ligament, tendon, etc.</td>
<td>Most abundant collagen type. Main constituent of major fibre bundles that give strength to connective tissues.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[α1(I)]3</td>
<td>Dermis and dentin, embryonic tissues</td>
<td>A minor form.</td>
</tr>
<tr>
<td>II</td>
<td>α1(II)</td>
<td>[α1(II)]3</td>
<td>Cartilage, intervertebral disc</td>
<td>Main cartilage collagen. Forms the main fibrils in this tissue.</td>
</tr>
<tr>
<td>III</td>
<td>α1(III)</td>
<td>[α1(III)]3</td>
<td>Blood vessels, dermis, intestine, etc.</td>
<td>Present in most tissues containing type I collagen but absent in bone and tendon.</td>
</tr>
<tr>
<td>V</td>
<td>α1(V)</td>
<td>[α1(V)]3</td>
<td>Widespread in low quantities: appears associated with collagen I fibrils</td>
<td>May form the core in type I collagen fibrils.</td>
</tr>
<tr>
<td></td>
<td>α2(V)</td>
<td>[α1(V)]2α2(V)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>α3(V)</td>
<td>α1(V)α2(V)α3(V)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XI</td>
<td>α1(XI)</td>
<td>α1(XI)α2(XI)α3(XI)</td>
<td>Cartilage, intervertebral disc</td>
<td>Forms fibrils which are associated with collagen II.</td>
</tr>
<tr>
<td></td>
<td>α2(XI)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>α3(XI)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Modified from Bateman et al. 1996)

2.1.2.1 Collagen types I, II and III

Type I collagen, the first member of the collagen protein family to be characterized, is the most abundant collagen in the human body, being found especially in bone, tendon, ligament and skin. It consists of two identical α1(I) chains and one different α2(I) chain. Low levels of an [α1(I)]3 homotrimer are also found in various tissues and cultured cells, and this has been found to possess higher amounts of 3-hydroxyproline and hydroxylysine than the corresponding heterotrimer (Jimenez et al. 1977, Moro & Smith 1977, Uitto 1979). Renaturation experiments with individual α chains of type I collagen have indicated that they are able to form both [α1(I)]2α2(I) heterotrimers and [α1(I)]3 homotrimers, although the former was favoured and the latter had a slightly lower Tm
Homotrimers with the structure of \([\alpha 2(I)]_3\) were also obtained at low temperatures, but their yield was much lower and their T_m was only about 20-24°C (Tkocz & Kühn 1969).

Fibrils of type I collagen often have large cross-sectional diameters and lengths and are often packed in parallel bundles, all features that give the tissues a high mechanical strength (see Kielty et al. 1993). In some cases the N propeptides are retained in the type I collagen molecules present in fibrils, although their removal is generally regarded as a prerequisite for proper fibrillogenesis (see Kühn 1987, Vuorio & de Crombrugghe 1990 for reviews). Immunochemical studies have shown that such molecules, called type I pN-collagen molecules, can be found on the surface of thin type I collagen fibrils. The retention of the N propeptide appears to be transient, since it is cleaved when the fibrils reach a diameter of 35-45 nm. It has therefore been suggested that the type I pN-collagen molecules regulate the fibril diameter (Miyahara et al. 1984, Fleischmajer et al. 1985, Fleischmajer 1986, Fleischmajer et al. 1988). The presence of type I pN-collagen molecules varies between tissues, and they are found more abundantly during embryonic development (Fleischmajer et al. 1988). It has also been suggested that the N propeptides of type I procollagen, or shorter peptides derived from them, may regulate the production of type I collagen by a feedback mechanism (Hörlein et al. 1981, Schlumberger et al. 1988).

Type II collagen is a homotrimer consisting of three identical \(\alpha 1(II)\) chains and is the major collagenous component of cartilage (see Miller & Matukas 1969, Miller & Gay 1987). Besides its expression in cartilage, the vitreous body of the eye, the intervertebral disc and the inner ear, it can also be found in many non-condrogenic tissues during development (Brewton & Mayne 1992). Type II collagen account for 90-95% of all the collagens in cartilage (Eyre 1991) and is its most important structural component (Mendler et al. 1989).

The gene sequence coding for the N propeptide has an alternatively spliced exon that codes for a 69-amino acid cysteine-rich domain (Ryan & Sandell 1990). Type II procollagen molecules that either contain this cysteine-rich globular domain (IIA) or are lacking in it (IIB) have distinct distributions during the various stages of condrogenesis, type IIA predominating in the precondrogenic mesenchyme and differentiating chondrocytes, while type IIB predominates in differentiated chondrocytes (Nah & Upholt 1991, Sandell et al. 1991). In cartilage the type II collagen molecules are polymerized to form collagen fibrils. Small-diameter fibrils (10-25 nm) are formed pericellularly, while much larger fibrils (up to 300 nm diameter) are formed in the territorial and interterritorial matrix. The type II collagen molecules in the fibrils overlap with each other by a distance of about a quarter of their length, thus forming a banded fibril. The molecules are covalently cross-linked between the triple-helical domains and the telopeptides (see Eyre et al. 1991, Kielty et al. 1993, Eyre & Wu 1995).

Type III collagen forms only homotrimeric molecules of three \(\alpha 1(III)\) chains. This collagen is typically found together with type I in many connective tissues (Miller et al. 1971, Miller & Gay 1987). It is the major collagen in blood vessels, but it is also found in the skin, lung, cornea and a number of other tissues. It is important for the development of the skin and the cardiovascular system and for the maintenance of the normal physiological functions of these organs in adult life (see Olsen 1995). It differs from the collagens of type I and type II in that it contains interchain disulphide bonds at the
extreme C-terminal end of the triple-helical region (Epstein, Jr. & Munderloh 1975). In addition, type III collagen contains a single cleavage site for trypsin in the triple-helical region (Miller et al. 1976). The fibrils are thinner than those of type I collagen and are especially prevalent in tissues exhibiting a high degree of elasticity, such as skin, aorta, gut and lung. Type III also forms heteropolymseric fibrils along with type I in many tissues that contain type I collagen, but is absent in bone and tendon (Keene et al. 1987a, Bornstein & Sage 1989). Type III collagen may also retain its N propeptide, yielding type III pN-collagen. Unlike the N propeptide of type I, which appears to have a transient existence on the fibril surface, the N propeptide of type III seems to remaining permanently on the surface of the fibrils (Fleischmajer et al. 1985).

2.1.2.2 Collagen types V and XI

The low abundance fibrillar collagen type V collagen is co-distributed with type I and is closely related to types I-III with respect to the length of the triple-helical domain (see Fichard et al. 1995). Type V molecules are characterized by a single, continuous triple helix with N and C terminal non-helical sequences, and pepsin digestion produces rod-like molecules of approximately 300 nm. Three genetically distinct type V collagen α chains, α1(V), α2(V) and α3(V), form homotrimers and heterotrimers with variable chain compositions: [α1(V)]3, [1(V)]2α2(V) and α1(V)α2(V)α3(V) (see Eyre et al. 1987). The proα1(V) chain is unique among the chains of fibril-forming collagens, as it lacks potential cross-linking sites in both the N and C telopeptides and as its potential N-linked carbohydrate attachment sites in the N propeptide are not utilized (Greenspan et al. 1991). Type V collagen shares a number of characteristics with type XI. It is a minor constituent in a number of tissues and it associates with the much more abundant type I collagen to form heterotypic fibres and regulates the diameter of these fibres (Birk et al. 1988).

Type XI collagen is a quantitatively minor fibril-forming collagen to be found in cartilage. Like type II it is also expressed in the vitreous body of the eye and the nucleus pulposus of the intervertebral discs (Eyre et al. 1987, Mayne & Brewton 1993). As a fibril-forming collagen, it has a long, uninterrupted collagenous domain and globular domains at the ends of the molecule. The major triple helical domain (COL1) contains about 1000 amino acids and is flanked by a non-collagenous C terminal region (NC1) and by an N terminal region that consists of a minor triple-helical domain (COL2) flanked by non-collagenous domains (NC2 and NC3). Type XI collagen is a heterotrimer of three α chains: α1(XI), α2(XI) and α3(XI), of which the α1(XI) and α2(XI) chains are structurally related to the α1(V) and α2(V) chains, respectively (Burgeson et al. 1982). The α3(XI) chain is encoded by the type II collagen gene COL2A1, the only difference being that the α3(XI) chain undergoes more extensive post-translational modification (more glycosylation) (Burgeson et al. 1982). Type XI collagen is thought to associate and interact with the more abundant type II collagen in a manner analogous to the interaction between types V and I in other tissues (Morris & Bächinger 1987, Mendler et al. 1989). The function of type XI collagen is not clear, but it has been suggested that it may serve
both as a core for growing type II collagen fibrils and as a means of regulating the

Collagens V and XI resemble each other closely in their structural and biological
properties, and their α chains can also substitute for each other, so that a fraction of the
α1(XI) chains in cartilage molecules can be replaced by α1(V) chains (see Eyre et al.
1987), and a fraction of the α1(V) chains in bone can be replaced by α1(XI) chains
(Niyibizi & Eyre 1989). Furthermore, the ocular vitreous body contains α1(XI) and
α2(V) chains but no α2(XI) chain (Mayne et al. 1993). These findings have led to the
suggestion that collagens V and XI should be classified as a single type, collagen V/XI
(Fichard et al. 1995).

### 2.2 Biosynthesis of collagen

The biosynthesis of collagen is a complex process involving a number of co-translational
and post-translational modifications catalyzed by at least nine enzymes, of which many
are unique to collagens and collagen-like molecules. In addition, collagen assembly and
folding are assisted by many molecular chaperones. The fibril-forming collagens are
synthesized as large precursor molecules known as procollagens (Mₑ about 150 kDa) and
their biosynthesis can be divided into intracellular and extracellular events. The complex
intracellular biosynthetic pathways and extracellular processing lead to the formation of
stable fibres in the extracellular matrix (see Kielty et al. 1993, Kivirikko 1993, Prockop
Bateman 1999).

The intracellular steps in collagen biosynthesis begin with translation of the mRNA
coding for the polypeptide chains. These chains are synthesized on membrane-bound
ribosomes, after which they pass through the membranes into the lumen of the
endoplasmic reticulum (ER). The translation products contain an N-terminal signal
sequence, which targets the nascent procα chain into the ER. Intracellular processing starts
with cleavage of the signal peptide during or shortly after translocation across the
membrane of the rough ER by a signal peptidase. The hydroxylations of proline and
lysine residues begin before the translation is complete and continue as post-translational
modifications after the release of the complete polypeptide chains from the ribosomes
until triple helix formation of the newly synthesized polypeptide chains prevents any
further hydroxylation (Kivirikko et al. 1992, Kielty et al. 1993, Bateman et al. 1996,
Kivirikko & Pihlajaniemi 1998, Lamande & Bateman 1999). The hydroxylations of
proline and lysine residues are catalyzed by three enzymes, prolyl 4-hydroxylase (EC
1.14.11.2), prolyl 3-hydroxylase (EC 1.14.11.7) and lysyl hydroxylase (EC 1.14.11.4)
(Kivirikko & Pihlajaniemi 1998). After the hydroxylation of lysine residues, some of the
hydroxylysines are galactosylated by hydroxylysyl galactosyltransferase and some of the
galactosylhydroxylysine residues are further glucosylated by galactosylhydroxylysyl
glucosyltransferase. In addition, certain asparagine residues in the propeptides are
glycosylated (Kivirikko & Myllylä 1979).
After the intracellular processing, the procollagen molecules are secreted from the cells into the extracellular space, where the N and C propeptides are cleaved by specific proteinases. Finally, the collagen molecules self-assemble into fibrils which are stabilized by cross-links, the formation of which is catalyzed by lysyl oxidase. The formation of a triple-helical procollagen begins by association of the C propeptides of the three proα chains, which then fold into a triple helix beginning from the C-terminus and proceeding to the N-terminus (Kadler et al. 1990, Kielty et al. 1993, Prockop & Kivirikko 1995, Bateman et al. 1996).

### 2.2.1 Hydroxylation of proline residues

Prolyl 4-hydroxylase catalyses the hydroxylation of proline residues to 4-hydroxyprolines at the Y positions of the repeating -Gly-X-Y- triplets (see Kivirikko 1993, Kielty et al. 1993, Prockop & Kivirikko 1995, Bateman et al. 1996, Kivirikko & Pihlajaniemi 1998, Kivirikko & Myllyharju 1998, Lamande & Bateman 1999). The polypeptide chains of the fibril-forming collagens vary in their hydroxyproline content, in the range 92-126 residues per 1000 amino acids (Table 2). In all these collagens approximately 50 % of the proline residues are hydroxylated.

The function of the hydroxy group in the 4-hydroxyproline residues is to stabilize the collagen triple helix at physiological temperatures. Non-hydroxylated collagen α chains can fold into a triple helix at low temperatures, but the midpoint of the thermal transition (T_m) from helix to coil in the case of non-hydroxylated type I procollagen molecules is only 24°C (Berg & Prockop 1973, Rosenbloom et al. 1973, Kivirikko & Pihlajaniemi 1998), which is about 15°C lower than that for molecules consisting of fully hydroxylated polypeptide chains.

**Table 2. Amounts of 4-hydroxyproline and hydroxylysine in the polypeptide chains of the fibril-forming collagens.**

<table>
<thead>
<tr>
<th>Collagen type and chain</th>
<th>4-Hyp</th>
<th>4-Hyp+Pro</th>
<th>4-Hyp %</th>
<th>Hyl</th>
<th>Hyl+Lys</th>
<th>Hyl %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iα1</td>
<td>114</td>
<td>232</td>
<td>49</td>
<td>10</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>Iα2</td>
<td>105</td>
<td>219</td>
<td>48</td>
<td>12</td>
<td>32</td>
<td>38</td>
</tr>
<tr>
<td>IIα1</td>
<td>96</td>
<td>202</td>
<td>48</td>
<td>18</td>
<td>38</td>
<td>47</td>
</tr>
<tr>
<td>IIIα1</td>
<td>126</td>
<td>233</td>
<td>54</td>
<td>6</td>
<td>35</td>
<td>17</td>
</tr>
<tr>
<td>Vα1</td>
<td>109</td>
<td>227</td>
<td>48</td>
<td>35</td>
<td>55</td>
<td>64</td>
</tr>
<tr>
<td>Vα2</td>
<td>109</td>
<td>206</td>
<td>53</td>
<td>24</td>
<td>42</td>
<td>57</td>
</tr>
<tr>
<td>Vα3</td>
<td>92</td>
<td>191</td>
<td>48</td>
<td>43</td>
<td>58</td>
<td>74</td>
</tr>
<tr>
<td>XIα1</td>
<td>100</td>
<td>221</td>
<td>45</td>
<td>37</td>
<td>56</td>
<td>66</td>
</tr>
<tr>
<td>XIα2</td>
<td>96</td>
<td>208</td>
<td>46</td>
<td>40</td>
<td>57</td>
<td>70</td>
</tr>
<tr>
<td>XIα3</td>
<td>104</td>
<td>222</td>
<td>47</td>
<td>21</td>
<td>36</td>
<td>58</td>
</tr>
</tbody>
</table>

The values are given per 1000 amino acids.
The table is modified from that presented by Kivirikko et al. (1992)
Since the 4-hydroxyproline residues together with water molecules stabilize the molecular structure of the collagens via hydrogen bonds (see Fig. 2.), non-hydroxylated collagen polypeptide chains cannot form triple-helical molecules in vivo, and almost complete 4-hydroxylation of proline residues in the Y positions of the -X-Y-Gly- triplets is required for the formation of a molecule that is stable at 37°C (see Kivirikko & Myllylä 1982, Kivirikko et al. 1992, Kivirikko & Pihlajaniemi 1998).

A few proline residues in the X-positions of -Gly-X-Y- triplets are hydroxylated to 3-hydroxyproline by prolyl 3-hydroxylase, but only after 4-hydroxyproline has been formed in the Y-positions (see Kielty et al. 1993, Kivirikko 1993, Prockop & Kivirikko 1995, Kivirikko & Pihlajaniemi 1998). The functions of the 3-hydroxyproline residues are unknown at present and their amounts vary markedly between collagen types and between tissues within the same collagen type (Kivirikko & Myllylä 1982).

Fig. 2. Mechanism by which 4-hydroxyproline stabilizes the triple helix. Water-mediated hydrogen bonding links 4-hydroxyl groups belonging to 4-hydroxyproline with carbonyl groups in either glycine within the same chain or 4-hydroxyproline in the adjacent chain. Two chains of the triple helix are shown (Modified from Brodsky & Ramshaw 1997).
2.2.1.1 Prolyl 4-hydroxylase

Prolyl 4-hydroxylase (procollagen-proline, 2-oxoglutarate 4-dioxygenase, EC 1.14.11.2.) is a member of the group of 2-oxoglutarate dependent dioxygenases. It catalyzes the 4-hydroxylation of proline residues in collagens and other proteins containing collagen-like amino acid sequences (for reviews, see Kivirikko et al. 1989, Kivirikko et al. 1992). This enzyme has a crucial role in the synthesis of all collagens, since 4-hydroxyproline residues are essential for formation of the collagen triple helix at body temperature (see section 2.1.1). Prolyl 4-hydroxylase is located within the cisternae of the rough ER. The main biological substrates for vertebrate prolyl 4-hydroxylases are nascent collagen polypeptide chains containing repeated -Gly-X-Y- sequences. Prolines in the Y positions of these triplets are hydroxylated and adjacent amino acids influence the interaction of peptide substrates with the enzyme. The triple-helical conformation of collagenous peptides completely prevents hydroxylation (for reviews, see Kivirikko et al. 1989, Kivirikko & Myllyharju 1998, Kivirikko & Pihlajaniemi 1998).

The active prolyl 4-hydroxylase in vertebrates is an $\alpha_2\beta_2$ tetramer with a molecular weight of about 240 kDa and consisting of two types of inactive monomer. The $\alpha$ subunit of the main form of human prolyl 4-hydroxylase ($M_r$ about 64 kDa) consists of 517 amino acids (Helaakoski et al. 1989) and the $\beta$ subunit ($M_r$ about 57 kDa) of 491 amino acids (Pihlajaniemi et al. 1987), both being synthesized in a form containing a signal peptide of 17 additional residues. The $\beta$ subunit of prolyl 4-hydroxylase is identical to protein disulphide isomerase (see section 2.2.1.3). Two isoforms of the mouse and human $\alpha$ subunits have been isolated and characterized (Helaakoski et al. 1989, Annunen et al. 1997), so that the previously known $\alpha$ subunit is now called the $\alpha$(I) subunit and the recently isolated one the $\alpha$(II) subunit. Both isoforms form an active $\alpha_2\beta_2$ tetramer with the same $\beta$ subunit, one being called the type I enzyme, $[\alpha(I)]_2\beta_2$, and the other the type II enzyme, $[\alpha(II)]_2\beta_2$ (Helaakoski et al. 1995, Annunen et al. 1997). The type I enzyme is the main form in most cell types and tissues, while the type II enzymes is the main form in chondrocytes and capillary endothelial cells (Annunen et al. 1998). Neither of these $\alpha$ subunit isoforms contains the C-terminal sequence -Lys-Asp-Glu-Leu- (Helaakoski et al. 1989, Annunen et al. 1997), which has been shown to be required for the retention of many soluble proteins within the lumen of the ER (Pelham 1990). It has been demonstrated that the prolyl 4-hydroxylase tetramer is retained within the lumen of the ER by means of a retention signal present in its $\beta$ subunits (Vuori et al. 1992a).

The EC 1.14.11. group comprises fifteen enzymes with the same reaction mechanism, of which the most extensively studied are prolyl 4-hydroxylase, lysyl hydroxylase and $\gamma$-butyrobetaine hydroxylase. These enzymes are oxidoreductases that require Fe$^{2+}$, 2-oxoglutarate, O$_2$ and ascorbate. The 2-oxoglutarate is stoichiometrically decarboxylated during hydroxylation, with one atom of the O$_2$ molecule being incorporated into the succinate and the other into the hydroxy group formed on the proline residue (see Kivirikko & Myllyharju 1998, Kivirikko & Pihlajaniemi 1998). The $\alpha$ subunits of prolyl 4-hydroxylase contain the enzyme’s catalytic sites. Kinetic studies have indicated that the cosubstrates and the peptide substrate become bound to the enzyme in an ordered manner (see Fig. 3.), Fe$^{2+}$ being bound first, followed by 2-oxoglutarate, O$_2$ and then the peptide substrate. The reaction products are released in the reverse order, the hydroxylated
peptide first, followed by the CO$_2$ and succinate. Fe$^{2+}$ is not released between most catalytic cycles (Myllylä et al. 1977, Tuderman et al. 1977, Myllylä et al. 1978).

2.2.2 Hydroxylation of lysine residues and glycosylation of hydroxylysine residues

Lysyl hydroxylase catalyzes the hydroxylation of some of the lysine residues in -X-Lys-Gly- triplets, the resulting hydroxylysine residues acting as attachment sites for O-linked carbohydrate units and being essential for the formation of stable intermolecular cross-links in the collagen fibrils (for reviews, see Kivirikko et al. 1992, Kivirikko & Pihlajaniemi 1998). These cross-links are important for providing the collagen fibrils with tensile strength and mechanical stability (see Section 2.2.2). The hydroxylysine content of the polypeptide chains of the fibril-forming collagens varies markedly, only 17% of the lysine residues in the $\alpha_1$(III) chain being hydroxylated, as opposed to 70% of those in the $\alpha_2$(XI) chain (see Table 2). The amount of hydroxylysine also varies within the same collagen type in different tissues and even in the same tissue in different physiological and pathological conditions (see Kivirikko & Myllylä 1982). In addition, the extent of hydroxylation of lysine residues is higher in embryonic tissues than in adult tissues (Barnes et al. 1974, Ryhänen & Kivirikko 1974, Strawich & Glimcher 1983).

Two specific enzymes catalyze the formation of hydroxylysine-linked carbohydrate units. A galactosyltransferase (EC 2.4.1.50) catalyses the addition of a galactose unit to some of the hydroxylysine residues (monosaccharide), and a galactosylhydroxylysyl glucosyltransferase (EC 2.4.1.66) then catalyses the addition of glucose to some of the galactosylhydroxylysine residues (disaccharide). The structure of the disaccharide unit with its peptide attachment is 2-O-$\alpha$-D-glucopyranosyl-O-$\beta$-D-galactopyranosyl hydroxylysine, containing an unusual $\alpha$1$\rightarrow$2-O-glycoside bond between the glucose and the galactose. The amounts of the carbohydrate units in collagens also vary between collagen types and within the same collagen type from different tissues, and even the ratio of galactosylhydroxylysine to glucosylgalactosylhydroxylysine shows distinct
variation (Kivirikko & Myllylä 1979). The precise function of the hydroxylysine-linked carbohydrate units is not clear, but it has been suggested that they may regulate the lateral packing of collagen molecules within fibrils (see Kivirikko 1995). This suggestion is supported by the finding that there is a correlation between the carbohydrate content and fibril diameter. Experiments on the de novo generation of type I collagen fibrils from type I procollagen by enzymic cleavage of the propeptides have demonstrated that a collagen with higher amounts of hydroxylysine and glycosylated hydroxylysine residues forms thinner fibrils than the same protein with a normal degree of these modifications (Torre-Blanco et al. 1992, Kivirikko 1995).

Another suggestion for the functions of the hydroxylysine glycosylations is related to recent findings that collagens can directly serve as ligands for receptor tyrosine kinases, which are important factors in the control of cell growth, differentiation, metabolism and cell migration and in many other biological responses (Schlessinger 1997). Discoidin domain receptors 1 and 2 (DDR1 and DDR2) bind specifically to different types of collagen, the DDR1 receptor being activated by collagen types I, II, III and XI whereas the DDR2 receptor is activated mainly by types I and III (Shrivastava et al. 1997, Vogel et al. 1997). DDR2 receptor activation has been shown to be dependent on the glycosylation of the collagen molecules, since the capability of deglycosylated collagens to activate the DDR2 receptor decreases considerably, whereas deglycosylation did not affect the ability of the collagen to activate the DDR1 receptor (Vogel et al. 1997).

2.2.2.1 Lysyl hydroxylase

Lysyl hydroxylase (procollagen-lysine 2-oxoglutarate 5-dioxygenase, EC 1.14.11.4.) catalyzes the hydroxylation of lysine residues in -X-Lys-Gly- triplets of collagens and other proteins which contain collagenous sequences (for reviews, see Kivirikko et al. 1992, Kivirikko 1995). The active enzyme is a homodimer ($\alpha_2$) with a molecular weight of about 180 kDa (Turpeenniemi-Hujanen et al. 1980, Turpeenniemi-Hujanen et al. 1981, Myllylä et al. 1988). Lysyl hydroxylase is a glycoprotein containing asparagine-linked carbohydrate units, which have been shown to be critical for maximal catalytic activity (Turpeenniemi et al. 1977, Myllylä et al. 1988). Although lysyl hydroxylase has been localized to the ER, it contains no known ER retention signals such as KDEL (Myllylä et al. 1991, Hautala et al. 1992).

The mature lysyl hydroxylase polypeptide contains 709 amino acid residues after cleavage of a signal peptide of 18 residues (Hautala et al. 1992). The sequence of lysyl hydroxylase shows no significant similarity to those of the $\alpha$ subunits of prolyl 4-hydroxylase despite marked similarities in the catalytic properties of these enzymes (Myllylä et al. 1991). cDNAs have so far been cloned for three human lysyl hydroxylase isoenzymes, some of the properties of which have been characterized (Hautala et al. 1992, Valtavaara et al. 1997, Passoja et al. 1998). Lysyl hydroxylase 1 is expressed in a number of tissues (Heikkinen et al. 1994, Valtavaara et al. 1998) whereas expression of the recently identified isoforms may be more strictly regulated. An alternatively spliced variant, lysyl hydroxylase 2a, is highly expressed in the heart, placenta and pancreas.
(Valtavaara et al. 1997), while expression of lysyl hydroxylase 2b is highest in the heart and skeletal muscle. Lysyl hydroxylase 3 is most highly expressed in the heart, placenta and pancreas (Valtavaara et al. 1998, Passoja et al. 1998).

The catalytic properties of lysyl hydroxylase are very similar to those of prolyl 4-hydroxylase. They both require Fe$^{3+}$, 2-oxoglutarate, O$_2$ and ascorbate and have similar reaction mechanisms (see Kivirikko et al. 1992).

### 2.2.3 Procollagen assembly and secretion

The formation of a triple-helical procollagen molecule begins by association of the C propeptides of the three proα chains (Bächinger et al. 1981, Bulleid et al. 1996). The N and C propeptides first become folded and intrachain disulphide bonds are formed within them (Bächinger et al. 1981, Doerge & Fessler 1986). The proα chains then associate through non-covalent interactions within the folded C propeptides, and interchain disulphide bonds are formed between the C propeptides (Olsen et al. 1976). Formation of the disulphide bonds is catalyzed by protein disulphide isomerase (PDI, see section 2.2.3.1), which promotes rapid formation of the correct bonds (Freedman 1989, Noiva & Lennarz 1992). The C propeptides appear to have specific recognition signals that ensure type-specific assembly of individual procollagen chains (see McLaughlin & Bulleid 1998). The proα chains are then folded into the triple helix, beginning from the nucleus formed within the C-terminal region in a zipper-like manner and proceeding to the N-terminus (see Engel & Prockop 1991). The rate-limiting steps in triple helix formation are the cis-trans isomerizations of the proline peptide bonds, which have been shown to be accelerated by enzymes known as peptidyl-prolyl cis-trans isomerases (PPIs) (Lang et al. 1987).

The intracellular location, binding properties and expression characteristics of the heat shock protein Hsp47, a molecular chaperone residing in the lumen of the endoplasmic reticulum, have led to the proposal that it may be a collagen-specific chaperone (Nakai et al. 1992, Nagata 1996, Nagata 1998). Molecular chaperones do not modify polypeptides, but bind to hydrophobic regions of unfolded proteins, preventing aggregation, while the polypeptide is still being synthesized (for reviews, see Gething & Sambrook 1992, Becker & Craig 1994, Hartl et al. 1994). Hsp47 is structurally similar to serine protease inhibitors, although it has no protease inhibition activity (Nagata 1996). Coimmunoprecipitation experiments have demonstrated that procollagen molecules are associated with Hsp47 in the ER (Nakai et al. 1992) and are dissociated from it in the cis-Golgi compartment (Satoh et al. 1996). Hsp47 associates transiently with procollagen and has been thought to play an important role in collagen triple helix formation and quality control under stress conditions (Satoh et al. 1996). The assembly of triple-helical procollagen molecules in insect cells (Lamberg et al. 1996, Myllyharju et al. 1997, Tomita et al. 1997) and yeast (Vuorela et al. 1997) does not require any recombinant Hsp47, however, and thus this chaperone does not seem to be essential, or else it may be replaced by others.
Chaperones also serve as folding quality controllers preventing the secretion of misfolded proteins and promoting their degradation. This has been seen in the case of certain osteogenesis imperfecta patients, for example, whose mutant type I collagen has been bound to the immunoglobulin heavy chain-binding protein BiP (Chessler & Byers 1992, Chessler et al. 1993). BiP is a 78 kDa protein located within the lumen of the ER and thought to play a general role in protein folding and oligomeric assembly (Haas & Wabl 1983). Another ER-resident stress-induced protein, GRP94, has likewise been shown in coprecipitation studies to be bound to collagen (Nakai et al. 1992). Recently it has also been demonstrated that PDI acts as a molecular chaperone during the assembly of procollagen chains (Wilson et al. 1998). It appears to interact specifically with the propeptides of monomeric type I procollagen chains and to prevent their premature assembly or aggregation (Wilson et al. 1998).

The transport of procollagen molecules from the ER to the Golgi complex is currently poorly understood. It has recently been shown that triple-helical procollagen molecules form large electron-dense aggregates in a cis-Golgi compartment of fibroblasts and that these aggregates then move across the Golgi stacks without leaving the lumen of the Golgi cisternae (Bonfanti et al. 1998). Transport through the Golgi complex appears to occur by virtue of progressive maturation of the Golgi cisternae (Mironov et al. 1997). Gradual changes in the luminal environment are likely to lead to increased condensation of the procollagen aggregates, resulting in the formation of secretory granules, which are then subject to exocytosis (Bonfanti et al. 1998). If triple helix formation is prevented, the random coil polypeptide chains first accumulate within the cisternae of the rough endoplasmic reticulum and are then in part degraded and in part secreted at a delayed rate (see Kivirikko et al. 1992).

### 2.2.3.1 Protein disulphide isomerase (PDI)

The β subunit of prolyl 4-hydroxylase is identical to protein disulphide isomerase (PDI, EC 5.3.4.1.), a polypeptide that has a modular structure comprising the domains a, b, b’, a’ and c (Edman et al. 1985, Freedman et al. 1994, Darby et al. 1996, Kemmink et al. 1997, Freedman et al. 1998). The domains a and a’ are homologous and bear sequential and structural similarity to thioredoxin (Kemmink et al. 1995, Kemmink et al. 1997) which is a small, well-characterized ubiquitous cytoplasmic reducing protein (Buchanan et al. 1994, Holmgren 1995, Martin 1995). These two domains each contain the four amino acid sequence -Cys-Gly-His-Cys- responsible for the disulphide isomerase activity of the polypeptide (Vuori et al. 1992a, Lyles & Gilbert 1994, Darby & Creighton 1995). The structures of the b and b’ domains are also similar to that of thioredoxin, but they show no amino acid sequence similarity to it and contain no -Lys-Gly-His-Cys- motif (Kemmink et al. 1997, Kemmink et al. 1999). The c domain, which forms the acidic C terminus of the polypeptide, is assumed to be involved in the binding of calcium ions (Lebeche et al. 1994). In addition, there is a -KDEL- sequence at the end of the c domain which is needed for the retention of the PDI polypeptide within the lumen of the ER (Mazzarella et al. 1990, Pelham 1990).
PDI has been cloned from human, vertebrate and invertebrate species, several plants, micro-organisms, yeast and fungi (see Kivirikko & Pihlajaniemi 1998). The processed human PDI polypeptide consists of 491 amino acids (Pihlajaniemi et al. 1987).

PDI is a multifunctional protein (see Noiva & Lennarz 1992, Noiva 1999) which catalyzes the formation, rearrangement and breakage of disulphide bonds within several secretory and cell surface proteins (see Noiva & Lennarz 1992, Freedman et al. 1994, Freedman et al. 1995). In addition, it has peptide binding activity (Noiva et al. 1993), acts as a chaperone-like polypeptide and assists in the folding of various proteins (LaMantia & Lennarz 1993, Puig et al. 1994, Cai et al. 1994, Otsu et al. 1994, Hayano et al. 1995, Yao et al. 1997). PDI also serves a β subunit in the vertebrate prolyl 4-hydroxylase tetramer (see section 2.2.1.1) and the microsomal triacylglycerol transfer protein (MTP) αβ dimer, which is required for the assembly of very low density lipoproteins in the liver (Wetterau et al. 1990). An additional reported property is that of serving as a major cellular thyroid hormone-binding protein, but the physiological significance of this function is unknown (Yamauchi et al. 1987, Cheng et al. 1987).

2.2.4 Procollagen processing and fibril assembly

The N and C propeptides of procollagens must be cleaved by specific proteinases in the extracellular space before the molecules can self-assemble into fibrils (see Prockop & Hulmes 1994). Removal of the propeptides requires two enzymes, procollagen N proteinase (EC 3.4.24.14.) and procollagen C proteinase (EC 3.4.24.). The procollagen proteinases are endopeptidases that operate at neutral pH and require a divalent cation such as Ca²⁺ for maximal activity (see Hojima et al. 1989, Prockop et al. 1998).

There are two types of procollagen N proteinases. Type I cleaves the N propeptides from both type I and type II procollagens, whereas cleavage of the N propeptides from type III procollagen requires a separate enzyme (see Prockop et al. 1998). The C propeptides of procollagens I-III are all cleaved by a single procollagen C proteinase (see Kivirikko 1995).

The bond cleaved by the type I N proteinase in a number of collagens is either Pro-Gln or Ala-Gln. The type III N proteinase likewise cleaves a Pro-Gln bond, while the C proteinase cleaves at least Ala-Asp, Gly-Asp and Arg-Asp bonds. The procollagen N proteinases must have a triple-helical conformation in their substrates, whereas the C proteinase has been reported to act on both native and denatured procollagens (for reviews, see Kivirikko & Myllylä 1985, Olsen 1991, Kivirikko 1995). During biosynthesis of type I procollagen, the N propeptides are removed before the C propeptides, while in the case of type II procollagen the same processing occurs in the reverse order. Cleavage of the N propeptide from type III collagen is a very slow process, which explains why collagen fibres in many tissues contain significant amounts of partially processed type III molecules with intact N propeptides (Kivirikko 1995, Prockop et al. 1998).

Cleavage of the C propeptides reduces the solubility of the protein, and the collagen molecules are then capable of self-assembling into fibrils (Kadler et al. 1987, Prockop &
This process does not require any enzymic catalysis or other factors and will occur easily in collagen solutions in vitro. Collagen fibrillogenesis is discussed in section 2.2.4.1. After assembly of the collagen molecules into fibrils, covalent cross-links are formed that provide the fibrils with their tensile strength and mechanical stability. The cross-links are formed from lysine and hydroxylysine-derived aldehydes that are formed in a reaction catalyzed by lysyl oxidase, EC 1.4.3.13 (see Kiely et al. 1993, Prockop & Kivirikko 1995). This is a copper enzyme that catalyzes oxidative deamination of the ε-amino group in certain lysine and hydroxylysine residues, the corresponding aldehydes then being used for the formation of various cross-links (Kagan & Trackman 1991). The aldehydes formed in the lysyl oxidase reaction can form two types of collagen cross-link, either by aldol condensation between two aldehydes or by condensation between one aldehyde and one ε-amino group of an unmodified lysine, hydroxylysine or glycosylated hydroxylysine residue. The cross-links formed from a hydroxylysine-derived aldehyde are much more stable than those formed from a lysine-derived aldehyde (see Kiely et al. 1993, Kivirikko 1995, Prockop & Kivirikko 1995).

### 2.2.4.1 Fibril formation

Collagens are the major proteins in connective tissues, and their fibrils provide the mechanical strength for the tissues and form a favourable matrix for cell attachment in vivo (Viidic 1996, Adachi et al. 1997). The assembly of collagen molecules into fibrils is an entropy-driven process similar to that occurring in other protein self-assembly systems such as those giving rise to microtubules, actin filaments and flagellae (for a review, see Kadler et al. 1987).

The collagen molecules assemble into the long D-periodic fibrils (D = 67 nm) that constitute the principal tensile element in the connective tissues. Collagen fibrils can have diameters within a broad range from 20 to 500 nm, depending on the tissue and age. The self-assembly of fibril-forming collagens has been studied for many years by warming and neutralizing solutions of collagen extracted from tissues with cold acidic buffers (see Veis & George 1994). A system has been developed for studying the de novo formation of fibrils by initiating the process through enzymatic cleavage of procollagen using purified N and C proteinases (Kadler et al. 1987). This system enabled pointed tips to be found that ran in the C to N terminal direction of the growing fibrils (Kadler et al. 1990, Prockop & Hulmes 1994). Charged amino acids have been reported to play an important role in the formation of segment-long spacing crystallite fibrils (Kobayashi et al. 1992, Veis & George 1994), whereas it has been suggested that hydrophobic interactions may be important for the stabilization of assembled collagen fibrils with a D period (Hofmann et al. 1978, Veis & George 1994). Several agents have been shown to either inhibit or accelerate fibril formation. Glucose and urea, for example, inhibit fibril formation, while low amounts of SDS stimulate it (Hayashi & Nagai 1972, Hayashi & Nagai 1973). Chemical modifications of collagen molecules also alter the features of fibril formation. Methylation of carboxylic groups with methanolic hydrochloric acid (Rauterberg & Kühn
1968), and succinylation of ε-amino groups with succinyl anhydride deprive collagen of its fibril formation activity at neutral pH (Hattori et al. 1999).

The collagens of types IX, XII, XIV, XVI and XIX are members of the subfamily of fibril-associated collagens with interrupted triple helices (FACITs) which are found at the surfaces of collagen fibrils and may participate in fibril formation (Pihlajaniemi & Rehn 1995, Bateman et al. 1996). Type IX collagen associates with the surfaces of the fibrils in the cartilage, vitreous body and developing cornea, where it is covalently bound to collagen types II and XI (Wu et al. 1992). It has been proposed, based on immunolocalization using monospecific antibodies, that type XI collagen forms a small cylindrical core that is surrounded by molecules of type II. Type IX collagen appears to be located exclusively on the fibril surface, where it may act as a molecular linker between collagen fibrils and macromolecules in the extracellular matrix. Types XII and XIV have very large N-terminal domains, and have been implicated in modulating the deformability of the extracellular matrix, as they have been localized near the surface of banded collagen fibrils and shown to mediate interactions between fibrils in vitro (Watt et al. 1992, Nishiyama et al. 1994).

2.3 Expression of recombinant proteins in a baculovirus system

The recombinant baculovirus expression system in insect cells is one of the most commonly used systems for the expression of foreign genes in eukaryotic cells. It is a helper-independent viral system which has been used to express large quantities of heterologous proteins from many sources, including mammals, fungi, plants, bacteria and viruses (Luckow & Summers 1988, O'Reilly et al. 1992, King & Possee 1992, Kidd & Emery 1993). It has many advantages over other expression systems. Baculoviruses are highly species-specific and are infectious only to arthropods and not to mammals or plants. In addition to high expression levels, the post-translational modifications of the foreign gene products are closely similar to those performed in mammalian cells (Luckow 1991). Furthermore, recombinant protein production is easy to scale up when suitable insect cell lines that grow in suspension cultures are used (van Lier et al. 1992).

The baculovirus most widely used for the expression of foreign genes in cultured insect cells is the Autographa californica Nuclear Polyhedrosis Virus (AcNPV) (Vialard et al. 1995). Although it was first isolated from the alfalfa looper (Autographa californica), it multiplies readily in cell lines derived from both the fall armyworm (Spodoptera frugiperda) and the cabbage looper (Trichoplusia ni). However, if the foreign gene is to be expressed in insect larvae rather than cultured insect cells, the silkworm baculovirus, Bombyx mori NPV (BmNPV), is a better choice (Maeda 1989).

The baculoviruses belong to a family of large, double-stranded DNA viruses and they can be divided into two subfamilies, the Eubaculoviridae (occluded baculoviruses) and the Nudibaculoviridae (non-occluded baculoviruses) (Couch 1991, Adams & McClintock 1991). The Eubaculoviridae produce crystalline proteinaceous structures called occlusion bodies, which are absent in the Nudibaculoviridae. The Eubaculoviridae subfamily is made up of two genera of granulosis and nuclear polyhedrosis viruses (NPV)
distinguished by the major protein that constitutes the occlusion bodies, the matrix granulin and polyhedrin, respectively. The diameter of nuclear polyhedrosis virus occlusion bodies is 1-15 µm and they are composed of the polyhedrin protein (Rohrmann 1986).

The AcNPV is the most extensively studied baculovirus strain, and its entire 128 kb genome has been sequenced (Harrap 1972, Kool & Vlak 1993, Ayres et al. 1994). The virus exhibits a biphasic life cycle both in the insect host and under tissue culture conditions. The first phase (10-24 h post-infection) involves the formation of enveloped virions that bud through the cellular membrane to form extracellular particles, known as budded viruses (BV). The BVs are infectious to neighbouring cells within the body of the host and in tissue culture, and spread the infection. The second phase (24-72 h post-infection) involves high-level expression of a few viral genes, in particular the polyhedrin gene. This 28 kDa protein is produced in such amounts that it can account for 50 % of the total protein content of an infected cell, and it is responsible for embedding the mature virus particles within the cell nuclei, generating very large viral occlusion bodies. These occluded viruses (OV) are released only after cell death and are essential for lateral transmission of the virus, protecting its particles from the external environment such as UV light, desiccation and nucleases (Kidd & Emery 1993, Vialard et al. 1995).

The baculovirus expression system is based on the introduction of a foreign gene into non-essential regions of the viral genome through allelic replacement. Production of the recombinant protein is achieved by infecting insect cells or larvae with the newly engineered virus. Several features of the baculovirus life cycle make this system open to genetic manipulation. Some genes, e.g. the polyhedrin gene, are unnecessary for the production of baculoviruses in cell culture (Smith et al. 1983) and can therefore be replaced with a foreign gene, which is positioned under control of the powerful polyhedrin promoter. Because the polyhedrin promoter is not activated until infectious virus particles are produced, the probability of expression of the foreign gene rather than the production of the virus is greatly minimized.

In order to obtain biologically active recombinant proteins, the choice of the right production system is particularly important. The recombinant baculovirus expression system has several unique features which are essential for many applications. This system typically produces recombinant proteins that are properly folded, disulphide bonded and oligomerized, and localized in the same subcellular compartment as the authentic protein (Kidd & Emery 1993). Insect cells are capable of performing several post-translational modifications, including N- and O-linked glycosylation, phosphorylation, acylation, amidation, carboxymethylation, isoprenylation, signal peptide cleavage and proteolytic cleavage. The sites where these modifications occur are often identical to those of the authentic protein in its native cellular environment (Hoss et al. 1990, Kloc et al. 1991, Kuroda et al. 1991). Thus it is possible to express a recombinant protein that is similar to its native counterpart both structurally and functionally in insect cells. However, the high expression level of the gene product of interest in this system may overwhelm the ability of the cell to modify the protein product, thus resulting in low levels of glycosylation or phosphorylation of the target protein. The highest expression level reported for recombinant proteins in insect cells is 50 % of the total cellular protein, which is approximately 1 g of recombinant protein per 1 × 10⁹ cells (Crossen & Gruenwald 1998).
Recombinant viruses can currently be produced easily at high frequency and there are a wide variety of vectors available for the expression of single gene products in non-secreted, secreted or membrane-bound forms, with or without affinity tags to aid protein purification (Crossen & Gruenwald 1998). Co-expression of protein-modifying enzymes using recombinant baculoviruses with multiple promoters, or using different recombinant baculoviruses simultaneously, can be used to enhance the production of functional recombinant proteins in insect cells (Lamberg et al. 1996, Myllyharju et al. 1997, Hollister et al. 1998, Kulakosky et al. 1998, Pajot-Augy et al. 1999). Recombinant baculovirus-infected insect cells also provide a useful system for studying the viral particle assembly process (Newcomb et al. 1996). This system is being increasingly utilized for the development of vaccine candidates based on the production of virus-like particles and conventional recombinant antigens (Goldmann et al. 1999). It has also been proposed that modified baculovirus expression vectors containing mammalian gene regulatory elements can be used for gene delivery and expression in mammalian cells in vivo (Hofmann et al. 1995, Condreay et al. 1999).

2.4 Expression of recombinant proteins in the yeast *Pichia pastoris*

About thirty years ago, Koichi Ogata described a new yeast species which can utilize methanol as its sole source of carbon and energy (Ogata et al. 1969). Veenhuis et al. (1983) showed that methanol utilization in *P. pastoris* requires a novel metabolic pathway involving several unique enzymes. The conceptual basis for the *Pichia pastoris* expression system stems from the observation that some of these enzymes are present at substantial levels only when the cells are grown on methanol (Egli et al. 1980, Veenhuis et al. 1983). The enzyme alcohol oxidase (AOX) catalyzes the first step in the methanol utilization pathway, namely the oxidation of methanol to formaldehyde and hydrogen peroxide. To avoid hydrogen peroxide toxicity, this first step takes place within a specialized organelle, the peroxisome, that sequesters the toxic hydrogen peroxide away from the rest of the cell. *P. pastoris* has two genes that encode alcohol oxidase, *AOX1* and *AOX2*. Expression of the gene for *AOX1*, which is responsible for the vast majority of such activity in the yeast cell (Ellis et al. 1985, Tschopp et al. 1987, Cregg et al. 1989), is tightly regulated at the level of transcription and appears to involve two mechanisms, a repression/derepression mechanism and an induction mechanism. Methanol is essential to induce high-level transcription of the *AOX1* gene, and the transcription is completely repressed by the presence of glucose or glycerol (Tschopp et al. 1987).

The molecular genetic manipulations of *P. pastoris*, such as DNA-mediated transformation, gene targeting, gene replacement and cloning by functional complementation (Cregg et al. 1985, Liu et al. 1992, Cregg & Russell 1998), are similar to those described for *Saccharomyces cerevisiae*, which was originally the model organism for the expression of heterologous proteins in yeast. The expression of a foreign gene in *P. pastoris* requires three basic steps, insertion of the gene into an expression vector, introduction of the expression vector into the *P. pastoris* genome, and examination of the potential expression of the foreign protein in the recombinant strains (see
Cereghino & Cregg 2000). In the P. pastoris system recombinant genes are commonly expressed under the control of the AOX1 promoter. Alternative promoters such as the P. pastoris GAP, FLD1, PEX8 and YPT1 promoters are also available when the use of methanol is not appropriate. The P. pastoris glyceraldehyde 3-phosphate dehydrogenase (GAP) gene promoter provides strong constitutive expression on glucose (Waterham et al. 1997), while the glutathione-dependent formaldehyde dehydrogenase (FLD1) gene can be highly induced by either methanol as a carbon source and ammonium sulphate as a nitrogen source or methylamine as a sole nitrogen source and glucose as a carbon source (Shen et al. 1998). Two moderately expressing P. pastoris promoters, PEX8 and YPT1, are also available. The PEX8 gene encodes a peroxisomal matrix protein and is expressed at low levels on glucose and induced to a modest extent when the cells are shifted to methanol (Liu et al. 1995). The YPT1 gene encodes a GTPase involved in secretion, and its promoter provides a low but constitutive level of expression in media containing either glucose, methanol or mannitol as a carbon source (Sears et al. 1998).

P. pastoris has no stable episomal vectors, and therefore the foreign DNA must be introduced into its genome by homologous recombination. As in S. cerevisiae, linear vector DNAs can be introduced and stable transformants can be generated via homologous recombination between the sequences shared by the vector and the P. pastoris genome. P. pastoris can be transformed by electroporation, a spheroplast generation method, or whole cell methods such as those involving lithium chloride and polyethylene glycol1000 (Cregg et al. 1985, Liu et al. 1992, Cregg & Russell 1998). Only a few selectable markers have so far been developed for the molecular genetic manipulation of P. pastoris. The existing markers are limited to the biosynthetic pathway gene HIS4 from either P. pastoris or S. cerevisiae, ARG4 from S. cerevisiae, and the Sh ble gene from Streptoalloteichus hindustanus, which confers resistance to the bleomycin-related drug zeocin (Cregg et al. 1985, Cregg & Madden 1989, Higgins et al. 1998). Recently, a new set of biosynthetic markers such as the genes for P. pastoris ADE1 (PR-amidoimidazole succinocarboxamide synthase), ARG4 (argininosuccinate lyase) and URA3 (orotidine 5’-phosphate decarboxylase) have been isolated, characterized and incorporated into expression vectors. In addition, a series of host strains containing all possible combinations of ade1, arg4, his4, and ura3 auxotrophies have been generated (see Cereghino & Cregg 2000).

High-level expression of a recombinant protein often requires the generation and isolation of a multicopy expression strain. In P. pastoris multiple gene insertion events at a single locus occur spontaneously at a low but detectable frequency. The use of multicopy transformants has been shown to markedly increase the yields of recombinant proteins (see Cereghino & Cregg 2000), but sometimes the resulting protein is more heterogeneous than the corresponding protein produced in a single-copy strain (Clare et al. 1991). There are three approaches to generating multicopy P. pastoris strains. The first involves the construction of a vector with multiple head-to-tail copies of the expression cassette, the second utilizes expression vectors that contain not only the P. pastoris HIS4 gene but also the bacterial Tn903 kan’ gene, which confers resistance to the related eukaryotic antibiotic G418 (Scorer et al. 1994), and the third involves the use of a vector with the bacterial Sh ble gene, which confers resistance to the antibiotic zeocin (Higgins et al. 1998). P. pastoris integrants are very stable in the absence of any selective pressure, even when present in multiple copies (see Cregg et al. 1993).
Foreign proteins expressed in *P. pastoris* can be produced either intracellularly or extracellularly. As this yeast secretes only low levels of endogenous proteins, the secreted heterologous protein constitutes the vast majority of the total protein in the medium. Several secretion signal sequences (such as *S. cerevisiae* α-mating factor prepro peptide or *P. pastoris* acid phosphatase *PHO1* signal peptide), including those present on heterologous proteins, have been used successfully for the secretion of recombinant proteins in *P. pastoris*. Recently, the signal peptide PHA-E from the plant lectin *Phaseolus vulgaris* agglutinin was found to be effective for the secreted expression of two plant lectins and the green fluorescent protein (Raemaekers *et al.* 1999). *P. pastoris* has a potential for performing many of the post-translational modifications typically associated with higher eukaryotes. These include processing of signal peptides (of both the pre and prepro-type), folding, disulphide bridge formation, certain types of lipid addition, and O- and N-linked glycosylation. Eukaryotic cells assemble O-linked saccharides onto the hydroxyl groups of serine and threonine. The O-linked oligosaccharides in mammals are composed of a variety of sugars, including N-acetylglalactosamine, galactose (Gal), and sialic acid, whereas lower eukaryotes such as *P. pastoris* add O-linked oligosaccharides composed solely of mannose (Man) residues. *P. pastoris* may glycosylate a heterologous protein even if that protein is not glycosylated by its native host. There is little information concerning the mechanism and specificity of O-glycosylation in *P. pastoris* (Goochee *et al.* 1991), but N-glycosylation in all eukaryotes begins in the ER, with transfer of a lipid-linked oligosaccharide unit Glc1,Man9,GlcNAc2 (Glc = glucose; GlcNAc = N-acetylglucosamine) to an asparagine in the recognition sequence Asn-X-Ser/Thr. This oligosaccharide core is then trimmed to Man8,GlcNAc2. At this point, the glycosylation patterns of lower eukaryotes (such as *P. Pastoris*) and begin to differ from those of the higher ones. The mammalian Golgi apparatus performs a series of trimming and addition reactions that generate oligosaccharides composed of Man5-6,GlcNAc2 (high-mannose type), a mixture of several sugars, or a combination of both (Goochee *et al.* 1991). In *S. cerevisiae*, N-linked core units are elongated in the Golgi through the addition of mannose outer chains, which are typically 50-150 mannose residues in length, a condition referred to as hyperglycosylation. Although N-linked high-mannose oligosaccharides added to proteins by yeast secretory systems represent a significant problem in the use of foreign-secreted protein, this hyperglycosylation is not usually seen in *P. pastoris*.

A hallmark of the *P. pastoris* system is the ease with which expression can be scaled up from shaker-flasks to high-density fermenter cultures, in which the expression levels of recombinant proteins are usually significantly increased because, firstly, environment it is possible to grow the organism to high cell densities in the controlled fermenter (>100 g/l dry cell weight or 500 OD600 U/l) (Higgins & Cregg 1998), and secondly, the level of transcription initiated from the *AOX1* promoter can be 3–5 times greater in *P. pastoris* cells fed with methanol at growth-limiting rates in fermenter culture than in cells grown in excess methanol in shaker flasks. Thirdly, methanol metabolism utilizes oxygen at a high rate and only in the controlled environment of a fermenter is it feasible to monitor and adjust the oxygen level in the culture medium accurately. In a fermenter, the strains are grown initially in a defined medium containing glycerol as the carbon source to accumulate biomass. When the glycerol is depleted, a transition phase is initiated in which additional glycerol is fed to the culture at a growth-limiting rate. Finally, methanol,
or a mixture of glycerol and methanol, is fed to the culture to induce expression (see review by Cereghino & Cregg 2000). Production of heterologous proteins in *P. pastoris* is ideal, because the medium components are inexpensive and defined, consisting of pure carbon sources (glycerol and methanol), biotin, salts, trace elements and water. In addition, as *P. pastoris* is cultured in media with a relatively low pH and methanol, it is less likely to become contaminated by most other microorganisms. The *P. pastoris* expression system has been used to produce a wide range of heterologous proteins up to ≥10 g/l, but there are also many examples of expression levels below the 1 g/l range (see Cereghino & Cregg 2000). In conclusion, the *P. pastoris* expression system is one of the most productive eukaryotic expression systems available.

2.5 Recombinant human collagens in therapeutic applications

The collagens make up the main protein component in all connective tissues, including skin, cartilage, bone, tendon, ligament and blood vessels. The various collagens have been shown to have complex roles in these tissues, and they are involved in a variety of important molecular interactions (for reviews, see Kivirikko 1993, Kielty *et al.* 1993, Pihlajaniemi & Rehn 1995, Prockop & Kivirikko 1995, Bateman *et al.* 1996). Because of the crucial role of collagens in tissues, it has long been assumed that they could potentially provide a basis for the biomaterials used in various fields of medicine. Our greater understanding of the molecular biology of collagens and their complex biosynthetic pathways now provides a broad range of new opportunities for collagen-based biomaterials (see Bateman *et al.* 1996, Pachence 1996).

There are two distinct groups of collagen-based biomaterials, tissue-based biomaterials, where the intact tissue is adapted so that the product meets the medical needs, and purified collagens, either as a powdered product or as a soluble product, which can then be reconstituted to a particular shape or form. A broad range of medical applications exist for collagens today as biomaterials for soft tissue repair and also hard tissue repair. These applications include injectable collagens for soft tissue augmentation, dressings for healing wounds and burns, bone repair, and replacement components for the cardiovascular system, such as bioprosthetic heart valves. One important application of collagen in the form of sponges or fleeces is its use as a haemostatic agent in surgery, which takes advantage of the inherent platelet binding property that leads to an initiation of the clotting cascade (Pachence 1996, Ramshaw *et al.* 1996). Gelatin is a well known biopolymer and has a long history of use, mainly as a gelling agent in food. It is, in essence, denatured and partial hydrolysed collagen and is prepared by hot acid or alkaline extraction of animal tissues such as bone and hides (Asghar and Henrickson 1982). The gelling properties of gelatin are due to the remarkable amino acid sequence of its collagen ancestor. The denatured collagen triple helices in a heated gelatin preparation renature upon cooling to form randomly interwining triple helices, resulting in the characteristic gel formation. Currently gelatin is used in a wide variety of consumer and medical products range from candies and desserts to vaccines, drugs, medical devices, dietary supplements and cosmetics. The collagens used in these applications have been isolated from animal
tissues and are liable to cause allergic reactions in up to 3% of human subjects (Cooperman & Michaeli 1984). Furthermore, even if the patients do not experience any hypersensitivity reaction, some of them develop collagen antibodies in their serum (Siegle et al. 1984). Another problem in Europe since early of 1990s has been the cattle disease known as bovine spongiform encephalities (BSE). The infectious agent that cause BSE is prion (a proteinaceous particle) which is also known to induce the disease in sheep and coats. This has aroused a suspicion that prion may be able to cross between animal species and has hence raised the question of the suitability of bovine products, especially in medicine. The medical use of collagens has been limited to type I (used as a biomaterial and a delivery system for certain drugs), because only this collagen type is readily available from animal tissues. It is obvious, therefore, that an efficient large-scale recombinant expression system for the production of various human collagens would have many applications (Pachence 1996, Ramshaw et al. 1996).

The availability of recombinant human collagens would also allow the development of many new improved applications in the health sphere. The main problem in the production of recombinant collagens is that the expression and production of a functional protein is complicated by the necessary post-translational processing events, in particular the hydroxylation of proline residues (see section 2.2). Stable recombinant human type I and II collagens have been expressed in mammalian cells, but the expression levels were too low for large-scale production (Ala-Kokko et al. 1991, Geddis & Prockop 1993, Fertala et al. 1994). In these systems the recombinant collagens were hydroxylated by the endogenous prolyl 4-hydroxylase of the host cell. Attempts to produce recombinant collagens in high-level expression systems have shown that their endogenous prolyl 4-hydroxylase activity levels were too low to achieve sufficient hydroxylation (Tomita et al. 1995, Lamberg et al. 1996, Toman et al. 1999). Efficient expression of recombinant human collagens has been achieved using coexpression of collagen polypeptide chains with the two types of prolyl 4-hydroxylase subunit in insect cells and the yeast Pichia pastoris (Lamberg et al. 1996, Myllyharju et al. 1997, Vuorela et al. 1997). The folding of the procollagen chains into triple helical molecules occurred correctly in both systems, and the melting temperatures of the recombinant type I and III collagens produced in insect cells were equivalent to those of the native counterparts, and that of the type III collagen produced in the yeast was only slightly lower (Lamberg et al. 1996, Myllyharju et al. 1997, Vuorela et al. 1997).
3 Outlines of the present research

Collagens are used today commercially in numerous medical applications. The most important sources are bovine skin and bones, but these materials are liable to cause immunological reactions in human subjects and carry a risk of disease-causing contaminants. It is therefore obvious that recombinant human collagens would be good substitutes for the animal collagens.

Production of recombinant collagens presents major problems because of the post-translational modifications required to achieve a fully folded and stable triple-helical collagen molecule. Prolyl 4-hydroxylase and lysyl hydroxylase are enzymes required in collagen biosynthesis. In the absence of prolyl 4-hydroxylase activity, the collagen polypeptide chains do not form triple helices that are stable at body temperature, and in the absence of lysyl hydroxylase the lysines are not hydroxylated and the collagens are thus not glycosylated.

When this work began it had recently been demonstrated that a fully active recombinant human prolyl 4-hydroxylase tetramer can be produced in insect cells. In addition, it had been shown that recombinant human type III collagen molecules with stable triple helices and low amounts of hydroxylysine, or ones that are totally lacking in hydroxylysine, can be produced in this systems using coexpression with prolyl 4-hydroxylase subunits. This suggested that it should be possible to produce type II collagen molecules with stable triple helices in insect cells. It also seemed possible that coexpression with lysyl hydroxylase might lead to an increase in the hydroxylysine content of the resulting type II collagen. The availability of type II collagens with low and high hydroxylysine content was expected to make it possible to study the roles of the hydroxylation of lysine and the subsequent glycosylation of hydroxylysine in the formation of type II collagen fibrils in vitro. It also seemed important to study whether a prolyl 4-hydroxylase coexpression strategy can be developed for the large-scale production of recombinant human collagen types I-III in the yeast *Pichia pastoris.*
The specific aims were:

1. to coexpress human cartilage-specific type II collagen with the $\alpha$ and $\beta$ subunits of human prolyl 4-hydroxylase in insect cells and to characterize the recombinant collagen,
2. to study the role of the hydroxylysine content in fibril formation in the recombinant type II collagen \textit{in vitro},
3. to express homotrimeric, heterotrimeric and modified type I collagens in the yeast \textit{Pichia pastoris} and to characterize the resulting collagens, and
4. to produce and characterize recombinant human type I-III collagens in the yeast \textit{Pichia pastoris} and to scale up the production in a 2-litre fermentor equipped with an O$_2$ supply system.
4 Materials and methods

Detailed descriptions of the materials and methods are presented in the original papers I-IV.

4.1 Expression of human type II collagen using a baculovirus expression system (I, II)

4.1.1 Construction of baculovirus transfer vectors and generation of recombinant viruses (I, II)

The baculovirus transfer vector pVLC2A1 coding for the protα1 chain of human type II procollagen was generated by digesting the full length cDNA in pGEM-7 vector (Ala-Kokko et al. 1991) with XbaI and DraI and ligating it into the XbaI-SmaI-digested baculovirus transfer vector pVL1392 (Invitrogen). The cDNA for human type II procollagen in this pGEM vector corresponds to a gene construct in which exon I codes for the protα1 chain of human type I procollagen and exons 2-54 for type II procollagen (Ala-Kokko et al. 1991). The recombinant baculovirus transfer vector was cotransfected into Spodoptera frugiperda Sf9 insect cells with a modified Autographa californica nuclear polyhedrosis virus DNA (BaculoGold, Pharmingen) by calcium phosphate transfection (Gruenwald & Heitz 1993). The resultant viral pools were collected 4 days later, amplified twice, plaque purified and used for recombinant protein production (Gruenwald & Heitz 1993). The resulting recombinant virus, termed rhproCII, was checked by PCR (Malitschek & Schartl 1991).

The construct pVLC3A1NP/C2A1 was generated by PCR, replacing the signal peptide and N propeptide of the pVLC2A1 construct with those of the human protα1(III) chain. The recombinant virus was termed rhproCIIIII. A double promoter construct coding for the α and β subunits of human prolyl 4-hydroxylase was generated by cloning the
cDNAs for the α subunit (Helaakoski et al. 1989) into the NotI site below the p10 promoter and that for the β subunit (Pihlajaniemi et al. 1987) into the BamHI site below the polyhedrin promoter of the transfer vector p2Bac (Invitrogen). A NotI site was generated in the α subunit cDNA 46 bp upstream of the translation initiation codon by PCR for the purpose of this cloning. This p2Bacαβ construct was cotransfected into Sf9 cells as above, resulting in the recombinant virus 4PHαβ.

4.1.2 Expression and analysis of recombinant proteins in insect cell cultures (I, II)

Insect cells (Spodoptera frugiperda Sf9 or High Five H5, Invitrogen) were cultured in TNM-FH medium (Sigma) supplemented with 10 % foetal bovine serum (Bioclear), either as monolayers or in suspension in shaker flasks at 27°C. For the production of recombinant proteins, the insect cells were seeded at a density of 6 × 10^5 cells/ml in monolayers or 1.5 × 10^6 cells/ml in suspension cultures. To produce recombinant human type II procollagen, the cells were infected with 2-3 of the viruses rhproCII, rhproCIINIII, 4PHαβ and LH, a virus coding for human lysyl hydroxylase (Pirskanen et al. 1996). RhproCII and rhproCIINIII were used in a 5-10-fold excess over 4PHαβ, and 4PHαβ and LH were used in a ratio of 1:1 or 2:1. Ascorbate phosphate, 80 µg/ml, (Yamamoto et al. 1992) was added to the culture medium daily.

The infected cells were harvested 24-96 h after infection, washed with a solution of 0.15 M NaCl and 0.02 M phosphate, pH 7.4, and homogenized in a 0.2 M NaCl, 0.1 % Triton X-100 and 0.05 M Tris buffer, pH 7.4. The cell homogenates were centrifuged at 10,000 × g for 20 min and the remaining insoluble material was solubilized in 1 % SDS. Aliquots of the Triton X-100 and SDS-soluble fractions were analyzed by denaturing SDS-PAGE. Aliquots of the supernatants of the cell homogenates and cell culture media were digested with pepsin at 22°C for 1 h and the thermal stability of the recombinant pepsin-resistant collagen was studied by digestion with a mixture of trypsin and chymotrypsin at temperatures of 27-42°C (Bruckner & Prockop 1981).

4.1.3 Purification and characterization of recombinant type II collagen (I, II)

Type II procollagen was expressed in High Five cells cultured either as monolayers or in suspension in shaker flasks. Unless otherwise mentioned, all the following steps were performed at 4°C. The cells were harvested 76 h after infection, homogenized in a buffer containing 0.1 % Triton X-100, centrifuged and the supernatant of the cell homogenate passed through a DEAE cellulose column (DE-52, Whatman) to remove nucleic acids. The flow-through fractions containing the type II procollagen were pooled and digested with pepsin. This step converted the type II procollagen to type II collagen and digested
most of the non-collagenous proteins. The recombinant type II collagen was precipitated by adding solid NaCl to a final concentration of 4 M and centrifuged at 16,000 × g for 1 h. The pellet was dissolved in a 0.5 M NaCl, 0.5 M urea and 0.05 M Tris buffer, pH 7.4. The type II collagen was separated from the remaining contaminants by gel filtration on a Sephacryl HR-500 gel filtration column (Pharmacia). The fractions containing the type II collagen were pooled, dialyzed against 0.1 M acetic acid and lyophilized.

Amino acid analysis of the purified type II collagen was performed on a Beckman system 6300 amino acid analyzer. The melting curves were determined in a Jasco J-500 spectropolarimeter equipped with a temperature-controlled quartz cell with a path length of 1 cm (Gilford). The thermal transition curves were recorded at a wavelength of 221 nm by raising the temperature linearly by 30 °C/h. Hydroxylysine and its glycosides were separated by cation exchange chromatography on Dowex 50W-X8, and a glucosylgalactosylhydroxylysine standard was used to quantify the glucosylgalactosylhydroxylysine and galactosylhydroxylysine (Tenni et al. 1984).

### 4.1.4 Fibril formation experiments (II)

Stock solutions for fibril formation studies were prepared by dissolving the collagen preparations in 0.05 % acetic acid followed by centrifugation. The collagen solution was adjusted to a final concentration of 200 µg/ml and checked with a Jasco J-500A spectropolarimeter. The collagen self-assembly conditions followed the method described by Williams et al. (1978). The type II collagen concentration was lowered to 100 µg/ml in 30 mM K₂PO₄ and 135 mM NaCl, pH 7.4. The samples were then transferred to a thermocontrolled quartz cuvette. Fibril formation was triggered by increasing the incubation temperature to 34°C. Optical density at 313 nm was monitored in steps of 2 min with a Perkin-Elmer Lambda2 photometer, each recording being limited to 900 min. After the experiment the sample was centrifuged, and the concentration of the supernatant was measured by spectroscopy to determine the amount of aggregated collagen.

For electron microscopy studies, aliquots of the assembly mixtures were prepared in parallel with the turbidity time assay. After 1000 min, samples of 4 × 3 µl were transferred to formvar –coated copper grids and the fibrils were allowed to settle for 30 min. The buffer was the drained cautiously with a filter paper and three washing steps were performed. The fibrils were stained with freshly prepared 1 % uranyl acetate, dissolved in distilled water for 2 min, washed 3 times more and dried. The grids were examined using a Zeiss EM 109 electron microscope. Fibril width was measured with an image analysis system (Optoquant, Lübeck, Germany) using an internal distance of 10 periods in the collagen banding pattern.
4.2 Expression of human collagen types I-III using a yeast expression system (*Pichia Pastoris*) (III, IV)

4.2.1 Construction of the *P. pastoris* expression plasmids and generation of the recombinant *P. pastoris* strains (III, IV)

In order to study the expression of recombinant procollagen homotrimers in the yeast *P. pastoris*, cDNAs for the α1 chains of type I, II and III procollagens were each cloned into the expression vector pPICZB (Invitrogen) under the control of the *P. pastoris* alcohol oxidase 1 promoter (*P AOX1*) (Vuorela et al. 1997, Vuorela et al. 1999). The vector includes the *Streptoalloteichus hindustanus* ble gene, which confers resistance to Zeocin and can be used as a selectable marker of transformation into the host *P. pastoris*. In order to express type I procollagen heterotrimers, the expression vectors pBLARG IX, pBLARG SX and pBLADE IX (obtained from Dr. James Cregg, Keck Graduate Institute of Applied Life Sciences, USA) were used. These expression vectors also allow cloning of recombinant cDNAs under the control of the *P. pastoris* *P AOX1* promoter and are designed for targeted integration into the *P. pastoris* genome. The construction of the expression plasmids for the production of type II and III procollagen homotrimers and type I collagen homotrimers and heterotrimers with and without the N propeptide is described in detail in papers III and IV. All expression plasmids were propagated in the *Escherichia coli* strain TOP10F' and the cells were grown in LB medium containing ampicillin (100 µg/ml) or in a low salt LB medium (Invitrogen) containing Zeocin (100 µg/ml). Recombinant strains were generated by the electroporation method as described in the Manual of Methods for Expression of Recombinant Proteins in *Pichia pastoris* (Invitrogen). The *P. pastoris* strains GS200 (*his4, arg4*), which have non-functional histidinol dehydrogenase and argininosuccinate lyase genes and yJC300 (*his4, arg4, ade1*), with an additional mutation in the PR-amidoimidazole succinocarboxamide synthase gene, were used as hosts for the expression of recombinant proteins. Stable transformants were generated by homologous recombination between the sequence shared by the linearized expression plasmid and the *P. pastoris* genome, and the methanol utilization phenotype of the recombinant strains was confirmed to be methanol utilization plus.

4.2.2 Expression and analysis of the recombinant prolyl 4-hydroxylase and type I, II and III procollagens (III, IV)

The *P. pastoris* cells were cultured as described in the Manual of Methods for Expression of Recombinant Proteins in *Pichia pastoris* (Invitrogen), with the following minor modifications. The cells were grown in 25 ml shaker flask cultures in a buffered glycerol complex medium (BMGY, pH 6.0) containing 1 g/l yeast extract and 2 g/l peptone. To
induce the expression of recombinant proteins, the cells were transferred to a buffered methanol complex medium (BMMY, pH 6.0), and methanol was added to a final concentration of 0.5 % every 12 h. The cells were harvested after 60 h of methanol induction at 30°C, washed once, suspended in a cold 5 % glycerol, 1 mM Pefabloc SC and 50 mM sodium phosphate buffer, pH 7.4, and broken with glass beads (0.5 mm diameter) by vortexing. The lysate was centrifuged at 10000 g for 30 min and total protein concentrations in the supernatants were determined using the Bio-Rad protein Assay (Bio-Rad). Aliquots of the soluble fractions of the cell lysates were analyzed by SDS-PAGE under reducing conditions followed by silver staining or Western blotting with polyclonal antibodies 1675 and 1669 (FibroGen) recognizing the N propeptide of human proα1(I) and the C propeptide of human proα2(I) chains, respectively, or a monoclonal antibody 95D1A recognizing the collagenous regions of various collagen chains (Snellman et al. 2000). The amounts of type I and III procollagen were analyzed by radioimmunoassays for the trimeric N propeptide (PINP-RIA) or C propeptide (PICP-RIA) of human type I procollagen and the N propeptide (PIIINP-RIA) of human type III procollagen (Farmos Diagnostica), respectively. Aliquots of the soluble extracts were treated with pepsin (150 µg/ml) at pH 2.0-2.5 for 1-2 h at 22°C and analyzed by SDS-PAGE followed by silver staining, and the thermal stability of the recombinant pepsin-resistant collagen was studied by digestion with a mixture of trypsin and chymotrypsin at increased temperatures (Bruckner & Prockop 1981). Prolyl 4-hydroxylase activity in the soluble extracts was measured by a method based on the hydroxylation-coupled decarboxylation of 2-oxo[1-14C]glutarate with (Pro-Pro-Gly)10 as the peptide substrate (Kivirikko & Myllylä 1982).

4.2.3 Fermentation of the recombinant P. pastoris strains (III, IV)

Fermentation was carried out according to Invitrogen guidelines in a Biostat CT bench-top fermentor (B. Braun) equipped with a 2-litre water-jacketed stainless steel vessel, microprocessor control of pH, dissolved oxygen, agitation, temperature and nutrient feed and electronic foam control. Fermentation basal salts medium (1 litre) containing 4 % v/v glycerol was sterilized for 30 min at 122°C. After that, 2.4 ml/l trace salts (Invitrogen, PTM1) and cells obtained from a 100 ml start-up culture in BMGY were added into the vessel aseptically. Fermentation proceeded in three phases. During initial glycerol batch phase, agitation was set to 800 rpm, temperature to 30°C, and pH to 5.0. This phase was continued until the glycerol was completely consumed (24 h). The second, glycerol feed phase was initiated by continuously feeding the culture with a medium containing 50 % (w/v) glycerol and 12 ml/l PTM1 trace salts at a feed rate of 18 ml/h/l until the amount of glycerol fed in was about 100-200 g/l and a wet cell weight of 100-200 g/l was achieved. The feeding was stopped for 1 h to ensure total depletion of glycerol. The third, methanol feed phase was started by feeding methanol containing 12 ml/l of PTM1 trace salts in at a rate of 2 g/h/l for 18 h, increasing the rate to 4 g/h/l for 3 h, to 5 g/h/l for 3h and finally to 6-7 g/h/l for the rest of the phase, which lasted five or six days in total, and collecting cell samples every day. The cells reached a density of 300-400 g/l (wet cell weight) on day
five or six and were harvested by centrifuging at 2500 × g for 10 min at 4°C and frozen at -20°C.

4.2.4 Characterization of recombinant proteins produced in the 2-litre fermentor (III, IV)

Yeast cell samples of 100 mg were resuspended in 200 µl of breaking buffer containing 0.2 M NaCl, 50 mM Tris pH 7.4 and 0.1 % TritonX-100, and equal volumes of glass beads (diameter 0.5 mm) were added. The mixture was vortexed for a total of 4 min, followed by 30 s mixing and 30 s on ice. The sample was centrifuged for 15 min at 13000 g, the supernatant was transferred to a fresh tube, and the total protein concentration was determined using the Bio Rad Protein assay (BioRad). Aliquots of the soluble fractions of the cell lysates were analyzed by SDS-PAGE followed by staining with Coomassie blue, and prolyl 4-hydroxylase activity was analyzed as given in section 4.2.2. Other aliquots of the soluble fractions and the cell culture medium samples were analyzed by radioimmunoassays for the trimeric N and C propeptides of human type I procollagen (Farmos Diagnostica) or the N propeptide of human type III procollagen. The soluble fractions were further digested with pepsin for 1 h at 22°C and analyzed by SDS-PAGE followed by silver staining.

4.2.5 Purification and characterization of the recombinant collagens produced in the 2-litre fermentor (III, IV)

Recombinant collagens were purified by breaking the cells in 0.1 M HCl through agitation with glass beads using a Mini-Beater (Biospec Products). The homogenate was filtered through a Buchner Funnel without the filter paper (Polypropylene, Nalgene), digested with a final concentration of 0.2 mg/ml of pepsin for 18 h at 4°C and centrifuged at 10000 g for 30 min. Collagen was precipitated by adding acetic acid to 0.5 M and solid NaCl to a final concentration of 3 M for collagen of types I and II or 1 M for type III. The sample was centrifuged as above and the pellets were dissolved in 0.1 M HCl. The collagens were reprecipitated from a neutral solution by adjusting the pH to 7.4 with 1 M Tris and adding NaCl to 3 M or 1 M as above, and incubated for 1 hour at 4°C. The samples were centrifuged as earlier and the pellets were redissolved in 0.1 M HCl and dialyzed against 0.1 M acetic acid. The dialysed samples were gel filtered on a Sephacryl S-500HR column in the ÄKTA explorer system (Amersham Pharmacia Biotech). Amino acid analysis of the purified type I collagen samples was performed in an Applied Biosystems 421A amino acid analyser and N-terminal sequencing in an Applied Biosystems 477A pulse-liquid protein sequencer. Fibril formation studies were carried out by the same method as described for type II collagen (see section 4.1.4).
5 Results

5.1 Expression of human type II procollagen in insect cells (I,II)

In order to study whether it is possible to produce triple-helical human type II procollagen in insect cells, a baculovirus transfer vector coding for the proα1 chains of type II procollagen was generated. The cDNA used corresponds to a construct in which the first exon codes for the proα1 chain of human type I procollagen and exons 2-54 code for human type II procollagen (Ala-Kokko et al. 1991). The first exon codes for the signal peptide and 12 amino acids of the 139-residue N propeptide. A double promoter construct coding for the α (Helaakoski et al. 1989) and β (Pihlajaniemi et al. 1987) subunits of human prolyl 4-hydroxylase was also prepared, and viruses expressing human proα1(II) chains and prolyl 4-hydroxylase were then generated. Although two isoforms of the human prolyl 4-hydroxylase α subunit have been identified (Helaakoski et al. 1989, Annunen et al. 1997), only the type I α subunit was used here, which will therefore simply be referred to as the α subunit in this work.

Expression of the prolyl 4-hydroxylase tetramer (αβ2) in the insect cells was analyzed by means of a prolyl 4-hydroxylase activity assay (Kivirikko & Myllylä 1982). The level of prolyl 4-hydroxylase activity obtained with this double promoter virus was about 30 % higher than that obtained using similar amounts of two separate viruses for the two types of subunit. The increased enzyme activity level is probably due to the fact that in the case of a single virus all the infected cells began to express both types of enzyme subunit at the same time, whereas in the case of two separate viruses some cells may express only one type of subunit.

In order to produce triple-helical human type II procollagen, High Five insect cells were coinfected with the viruses coding for the proα1(II) chains and prolyl 4-hydroxylase. The cells were harvested 72 h after infection and homogenized in a buffer containing Triton X-100. Aliquots of the Triton X-100-soluble samples were digested with pepsin at 22°C for 1 h to study whether the proα1(II) chains had been assembled into triple-helical procollagen molecules. The triple helix of collagens is resistant to proteolytic enzymes, while non-triple-helical proα chains and the propeptides of procollagens are digested. The pepsinized samples and the cell culture media were then
analyzed by SDS-PAGE under reducing conditions. Pepsin-resistant α1(II) chains were seen in the Coomassie blue-stained gels, but the level of type II collagen expression was distinctly lower than that of type III collagen expression seen in experiments using viruses coding for the proα1(III) chains and the two types of prolyl 4-hydroxylase subunit. Less than 15% of the total type II collagen produced in the High Five cells was found in the culture medium. This fraction was not studied further.

Attempts to increase the level of type II procollagen expression were made by replacing the sequence coding for the signal peptide and the whole N propeptide of the proα1(II) chain with those of the proα1(III) chain. A new virus, proCIINIII, was generated and the cells were infected and analyzed as above. The expression level obtained for type II collagen with this new virus was about 4-8 times that obtained with the original virus and was found to be about 50 mg/l, corresponding to about 70-80 mg/l of type II procollagen. Quantifications of the type II collagen expression levels were based on measurements of the purified collagen.

When the Triton X-100-soluble proteins of homogenates of insect cells infected with the original virus coding for the proα1(II) chains were analyzed without pepsin treatment, two weak bands were seen. The upper band corresponded to full-length proα chains and the lower band to pNα chains (chains from which the C propeptide had been cleaved). Two major bands were seen in samples from insect cells infected with the modified proCIINIII virus, but in this case the upper band corresponded to pNα chains and the lower band to fully processed α chains. In both cases all these bands were stained in Western blots by antibody 95DIA, which recognizes collagenous sequences (Snellman et al. 2000). In the case of the modified virus, the upper band was also stained by antibody PIIINP against the N propeptide of human type III procollagen. The rate of conversion of the procollagen chains varied between experiments, but the rate of N-terminal conversion of the modified chains was always much higher than that of the non-modified proα1(II) chains.

The thermal stability of the pepsin-treated recombinant type II collagen present in the crude Triton X-100 extracts was studied by digestion with pepsin at 22°C followed by digestion with a mixture of trypsin and chymotrypsin by increasing the temperature gradually to 42°C (Bruckner & Prockop 1981). T_m of the recombinant human type II collagen was found to be about 40°C.

5.1.1 Characterization of the recombinant human type II collagen (I, II)

Type II collagens expressed in insect cells were purified as described in Section 4.1.3 and amino acid analyses were performed on the purified samples. The amino acid composition agreed well with that reported for non-recombinant human type II collagen (Keene et al. 1987b) and the 4-hydroxyproline content was identical, whereas the hydroxylysine content was only about 17% of that in the non-recombinant protein, with wide variation between experiments. The galactosylhydroxylysine and glucosylgalactosylhydroxylysine contents were correspondingly low. The T_m of the purified recombinant type II collagen, as determined by CD analysis, was 40.0°C.
5.1.2 Expression and characterization of recombinant human type II collagen with increased hydroxylysine content (I, II)

In order to study, whether it is possible to increase the hydroxylysine content of the recombinant type II collagen, High Five cells were coinfected with three viruses, two of them coding for the proα1(II) chains and prolyl 4-hydroxylase as above and the third coding for lysyl hydroxylase (Pirskanen et al. 1996). The type II collagens were purified as described in Section 4.1.3, and amino acid analyses were performed on the purified samples. The amino acid composition of this recombinant type II collagen was very similar to that of the non-recombinant form, the hydroxylysine content, about 19 residues/1000 amino acids, being even slightly higher. The extent of glycosylation of this fully lysine hydroxylated type II collagen was also measured, and the increase in the hydroxylysine content was found to have led to a corresponding increase in glycosylated hydroxylysine residues from about 2 to 9 residues /1000 amino acids, which is about 90 % of that in the non-recombinant protein. Nevertheless, the extent of glucosylation of galactosylhydroxylysine residues in the recombinant type II collagen was lower than that in the non-recombinant collagen. Only about 30 % of the carbohydrate present in the fully lysine hydroxylated type II collagen was found in the form of glucosylgalactosylhydroxylysine, whereas half of the carbohydrate present in the non-recombinant collagen is present in this form.

The thermal stabilities of the type II collagens with the high and low hydroxylysine content were essentially identical, as the Tm obtained for the pepsin-treated high hydroxylysine collagen by digestion with the mixture of trypsin and chymotrypsin was about 40°C and that obtained for the corresponding purified collagen by CD spectrum analysis was 40.3°C.

5.1.3 Formation of type II collagen fibrils in vitro (II)

In order to study whether hydroxylysine and its glycosides play a role in fibril formation, the recombinant type II collagen preparations with low and high hydroxylysine content were digested with pepsin and their fibril formation studied. In vitro fibril formation in these two kinds of recombinant type II collagen was evaluated using an assay based on the increase in absorbance at 313 nm of a 100 µg/ml collagen solution at 34°C as a function of time. A marked difference in this respect was found between the recombinant type II collagens produced without lysyl hydroxylase and with lysyl hydroxylase. The maximum absorbance of the collagen with a low hydroxylysine content was reached within 5 min, whereas the absorbance of the collagen with a high hydroxylysine content increased for about 600 min. Thus the maximum absorbance/µg collagen incorporated into the fibrils of the low hydroxylysine collagen was about 6 times that observed with the high hydroxylysine collagen. The critical concentrations for fibril formation also differed markedly between the two types of recombinant type II collagen. The apparent
critical concentration for the low hydroxylysine collagen was less than 10 µg/ml, while the corresponding value for the high hydroxylysine collagen was about 70 µg/ml.

Electron microscopy of the fibrils formed by the high hydroxylysine collagen showed that they were typically very thin, with essentially no interfibril interaction or fibril aggregation, and only a few slightly thicker fibrils with sharp tips on both ends could also be seen. In contrast, the low hydroxylysine collagen formed thick fibrils on a background of thin ones. In order to study the distribution of the fibril diameters, a histogram of 200 measured fibril diameters was plotted against their relative frequency. The high hydroxylysine collagen showed the highest frequency of fibrils with a diameter of about 10 nm, whereas the low hydroxylysine collagen had the highest frequency of fibrils with a diameter of about 20 nm, and some fibrils had diameters exceeding 60 nm.

5.2 Expression of type I procollagen with and without its N propeptide in *Pichia pastoris* in shaker flasks (III)

To study whether it is possible to produce the human type I procollagen homotrimer in *P. pastoris*, yeast strains were generated that expressed human prolyl 4-hydroxylase subunits and proα1(I) chains. The cDNA for the proα1(I) chain was cloned into the expression vector pPICZB (Invitrogen) and transformed into a recombinant *P. pastoris* strain expressing human prolyl 4-hydroxylase subunits (Vuorela et al. 1997, Vuorela et al. 1999) in which the β subunit had the *S. cerevisiae* αMF pre-pro sequence.

The *P. pastoris* cells were cultured as described in Section 4.2.2, and the expression of recombinant proteins was induced by adding methanol to a final concentration of 0.5 % every 12 h. The cells were harvested and broken 60 h after induction and then analyzed for the presence of the type I procollagen homotrimer in the soluble fraction of the cell extract and the culture medium. These analyses showed that the majority of the procollagen produced remained inside the cells and only a small amount had been secreted into the culture medium.

The *P. pastoris* strain expressing the type I procollagen homotrimer was found to produce full-length proα1(I) chains that formed triple-helical molecules with collagen domains which were resistant to pepsin digestion. The thermal stability of the pepsin-digested type I collagen homotrimer was studied by means of a brief digestion with a mixture of trypsin and chymotrypsin at various temperatures (Bruckner & Prockop 1981). The Tm of the recombinant type I collagen homotrimer was found to be approximately 38°C. The mean level of expression of the type I procollagen homotrimer was 110 ng/100 µg total protein, as measured with a radioimmunoassay for the trimeric N propeptide.

To study whether it is possible to produce heterotrimeric human type I procollagen in *P. pastoris*, a four-gene expression system was used. The initial approach to producing heterotrimeric procollagen was to generate a new strain expressing prolyl 4-hydroxylase by transforming the constructs pBLARGIXα (complementing for arg4) and pPIC9PDI (Vuorela et al. 1997) (complementing for his4) into a yJC300 (his4, arg4, ade1) host strain. This was followed by transfer of the pBLADEIIXproα2(I) construct (complementing for ade1) into this strain, generating the strain Proα2(I), and after that
the pPICZBproα1(I) construct, generating the strain Proα1(I)+Proα2(I). An alternative strategy utilizing a double expression construct pAO815proα2(I)/proα1(I) (complementing for his4) was also developed. Another strain expressing prolyl 4-hydroxylase, derived from transfers of pBLADEIXα and pBLARGSXPDI into the yJC300 strain was used as a host for integration of the pAO815proα2(I)/proα1(I) construct, and the strain was named Proα2(I)/Proα1(I).

The recombinant strains were cultured, induced and harvested as above and aliquots of the soluble fraction of the cell extracts were analyzed by SDS-PAGE followed by Western blotting. Full-length proα1(I) chains were seen in immunoblots of the strains Proα1(I), Proα1(I)+Proα2(I) and Proα2(I)/Proα1(I), and full-length proα2(I) chains were seen in those of the Proα2(I), Proα1(I)+Proα2(I) and Proα2(I)/Proα1(I) strains. The presence of triple-helical procollagen molecules was examined by digesting other aliquots of the cell extracts with pepsin. A single pepsin-resistant polypeptide corresponding to the α1(I) chains was seen in the digested sample of the Proα1(I) strain in silver-stained SDS-PAGE, whereas no pepsin-resistant polypeptide was seen in the corresponding sample from the Proα2(I) strain. When both types of proα chain were coexpressed in the Proα1(I)+Proα2(I) or Proα2(I)/Proα1(I) strain, pepsin-resistant polypeptides corresponding to the α1(I) and α2(I) chains were detected in silver-stained SDS-PAGE with a ratio of 2 to 1. All the pepsin-resistant α2(I) chains seen in these samples must have been present in heterotrimeric molecules, and as the ratio of the α1(I) and α2(I) chains was 2 to 1, essentially all the α1(I) chains must likewise have been present in heterotrimers. The mean expression level of the type I procollagen heterotrimer, when measured with a radioimmunoassay for the trimeric N propeptide, was 60 ng/100 µg total protein. No difference in this expression level was seen between the Proα1(I)+Proα2(I) and Proα2(I)/Proα1(I) strains.

In order to study whether type I pCα chains (chains lacking their N propeptide) form triple-helical homotrimeric and heterotrimeric pCcollagen molecules, the strains PCα1(I) and PCα1(I)+PCα2(I) were generated by transferring the constructs pPICZBpCα1(I) and pBLADEIXpCα2(I) into a strain expressing prolyl 4-hydroxylase as described above for the Proα1(I) and Proα1(I)+Proα2(I) strains. The strains were cultured, induced and broken as above, and the soluble fractions of the cell extracts were analyzed by SDS-PAGE followed by Western blotting. PCα1(I) chains were seen in the immunoblots of the strains PCα1(I) and PCα1(I)+PCα2(I), and pCα2(I) chains were seen in those of the PCα1(I)+PCα2(I) strain. The type I pCcollagen chains became efficiently assembled into triple-helical pCcollagen molecules, since a polypeptide corresponding to the α1(I) chains was seen in the pepsin-digested samples of the PCα1(I) strain and two polypeptides with a 2 to 1 ratio corresponding to the α1(I) and α2(I) chains were seen in the PCα1(I)+PCα2(I) strain. The levels of expression of the type I pCcollagen molecules were measured with a radioimmunoassay for the trimeric C propeptide, and surprisingly, those of the recombinant type I pCcollagen homotrimers and heterotrimers were found to be 1.5-3 times those of the corresponding recombinant type I procollagen molecules.
5.3 Expression and characterization of type II and III procollagens in *Pichia pastoris* in shaker flasks (IV)

In order to study the expression of recombinant human type II and III procollagens in *P. pastoris*, yeast strains expressing human prolyl 4-hydroxylase subunits and type II or type III procollagen were generated by cloning separately the cDNAs for the proα1(II) and proα1(III) chains into the expression vector pPICZB (Invitrogen) and transferring these into a recombinant *P. pastoris* strain expressing human prolyl 4-hydroxylase subunits (Vuorela *et al.* 1997, Vuorela *et al.* 1999) as explained above for the type I procollagen homotrimer.

Analysis of the expression of the type II and III procollagens showed that the majority of these procollagens also remained inside the yeast cells and only small amounts had been secreted into the culture medium. The *P. pastoris* strains expressing these two procollagens likewise produced full-length proα chains that formed triple-helical molecules with collagen domains resistant to pepsin digestion. The folding and thermal stability of the pepsin-digested type II and III collagens were studied by means of a brief digestion with a mixture of trypsin and chymotrypsin at various temperatures (Bruckner & Prockop 1981). The T<sub>m</sub> values of both recombinant collagens were approximately 38°C, and the expression levels obtained in the shaker flask cultures were 5-15 mg/l.

5.4 Production of recombinant human type I-III collagens in a 2-litre fermentor and characterization of the recombinant proteins (III, IV)

Many previous studies have shown that shaker flask conditions are not optimal for protein production in *P. pastoris* due to insufficient O<sub>2</sub>, and marked increases in expression levels are usually obtained in fermentors (Romanos *et al.* 1992, Cregg *et al.* 1993).

Therefore, to increase the level of expression of the type I-III collagens in *Pichia pastoris* and to obtain fully hydroxylated type I-III collagens, the yeast strains were cultured in a 2-litre fermentor equipped with an O<sub>2</sub> supply system. The cells were harvested six days after induction. The levels of expression of the type I-III procollagen homotrimers, the type I procollagen heterotrimers and the type I pCcollagen homotrimers and heterotrimers ranged from 200 to 600 mg/l.

The recombinant collagens were purified by pepsin digestion, which converts procollagen or pCcollagen to collagen and digests most of the non-collagenous proteins. This was followed by selective salt precipitation and Sephacryl S-500HR gel filtration in the ÄKTA explorer system (Amersham Pharmacia Biotech). Amino acid analyses showed that the purified recombinant collagen homotrimers and heterotrimers were identical in 4-hydroxyproline content to the corresponding non-recombinant human proteins. The only difference in amino acid composition between the recombinant and non-recombinant collagens was the lack of hydroxylysine and corresponding increase in lysine content in the recombinant molecules.
N-terminal sequencing of the polypeptide chains of the recombinant collagen homotrimers and heterotrimeric showed that in most cases the pepsin digestion had removed several residues from the N terminus of the telopeptide domain, but in the case of the α2(I) chain only 1-2 residues had been removed. Interestingly, N-terminal sequencing of the α chains derived from the type I pCcollagen homotrimers and heterotrimeric indicated that in most cases pepsin digestion had removed no amino acids from the telopeptide domains of the α1(I) and α2(I) chains, and only minor populations of the chains had lost five residues or one residue, respectively.

Fibril formation in the case of the recombinant collagens was studied by incubating the collagen solutions in pH 7.4 at 34°C in vitro. All the recombinant collagens were found to form native-type fibrils that showed the banding pattern characteristic of collagen fibrils in electron microscopy. This indicates that the differences observed in the N-termini of the molecules do not influence the properties of the fibrils.
6 Discussion

Type I collagen is now used in many medical applications (Rao 1995, Pachence 1996, Ramshaw et al. 1996, Ruszczak & Schwarts 1999), but the material used in all these cases have been isolated from animal tissues, usually from bovine skin, and therefore represents proteins that are foreign to humans and can cause various immunological reactions. In addition, they carry a risk of disease-causing contaminants. Collagens isolated from human tissues would seem much more suitable, but their availability is limited and they also carry a risk of contaminants. Consequently, recombinant human collagens are expected to have numerous applications. They also have the advantage that their availability is not limited to type I, as it might be possible to produce any human collagen type in a recombinant system. Some of the other collagen types may be more suitable for certain applications than type I.

Recombinant DNA technology offers a powerful tool for producing foreign proteins in host cells of various kinds, ranging from bacteria to mammalian cells. However, expression of recombinant collagens is difficult to achieve because most systems do not have sufficient levels of activity of prolyl 4-hydroxylase, which is an absolute requirement for the formation of a stable, triple-helical collagen molecule. Bacteria and yeasts (Vuorela et al. 1997) have no prolyl 4-hydroxylase activity, and insect cells (Lamberg et al. 1996) and the mammary gland (John et al. 1999) have insufficient levels of this enzyme activity. As a result, the recombinant collagen polypeptide chains expressed in most systems will remain in the form of non-triple-helical, non-functional proteins, or, if the cells are grown at low temperatures, the chains may form molecules with unstable triple helices.

Recombinant collagens have now been expressed using mammalian, insect and yeast cells and transgenic animals and plants, whereas no report on collagen expression and assembly using a bacterial system has been published. Single procollagen genes have been expressed in mammalian cells to produce homotrimeric type I procollagen, for example (Geddis & Prockop 1993), or type II procollagen (Ala-Kokko et al. 1991, Fertala et al. 1994) or homotrimeric type V collagen (Fichard et al. 1997), but the expression levels have been far too low for large-scale production. In the case of the baculovirus system, the two types of subunit of prolyl 4-hydroxylase were coexpressed and shown to form a functional enzyme tetramer (Vuori et al. 1992b). Subsequently, type
I and III procollagens were produced and shown to be partially hydroxylated by the endogenous prolyl 4-hydroxylase of the insect cells and fully hydroxylated by the coexpressed recombinant enzyme (Tomita et al. 1995, Lamberg et al. 1996, Myllyharju et al. 1997, Tomita et al. 1997). More recently, the baculovirus system including coexpression with prolyl 4-hydroxylase has also been used for the production of human collagens of type IX (Pihlajamaa et al. 1999) and type XIII (Snellman et al. 2000). Homotrimeric type I procollagen and an engineered form of its homotrimer have also been expressed in the milk of transgenic mice, either without recombinant prolyl 4-hydroxylase (Toman et al. 1999) or with it (John et al. 1999). The yeast Pichia pastoris was first engineered to express prolyl 4-hydroxylase, and was subsequently shown to produce functional type III procollagen molecules when the gene for the proα1(III) was coexpressed (Vuorela et al. 1997, Vaughn et al. 1998, Keizer-Gunnink et al. 2000). A fragment of type III collagen has also been coexpressed with prolyl 4-hydroxylase in Saccharomyces cerevisiae (Vaughn et al. 1998). Even transgenic tobacco plants have recently been shown to produce recombinant human type I procollagen homotrimers, but the trimers had triple helices of low thermal stability, as recombinant prolyl 4-hydroxylase was not coexpressed with them (Ruggiero et al. 2000).

6.1 Expression of type II collagen in insect cells and characterization of the resulting protein

One of the aims of this work was to express fully hydroxylated cartilage-specific type II collagen molecules in insect cells and to characterize the resulting protein. As the need for recombinant prolyl 4-hydroxylase in collagen production in insect cells had been clearly demonstrated in the case of recombinant type III collagen (Lamberg et al. 1996), all the type II collagen experiments were performed in the presence of the recombinant enzyme. The data indicated that coexpression of the proα1(II) chain with human prolyl 4-hydroxylase led to the formation of type II collagen molecules with stable triple helices. The level of expression of the type II collagen was nevertheless lower than that of type III, and we therefore studied whether expression of the proα1(II) chain can be increased by replacing the sequences coding for its signal peptide and that of the N propeptide by those of type III procollagen. This modification increased the expression level about 4-8-fold and the highest expression levels obtained with this modified construct were even slightly higher than those obtained for type III collagen (Lamberg et al. 1996). The use of the modified construct increased the expression level but it did not influence the composition of the processed collagen, as the propeptides are cleaved during the pepsinization used in the purification procedure.

The type II procollagen molecules produced in insect cells were found to be processed variably as a function of time. The wild-type type II N propeptide is quite stable and it is believed to fold back on the triple-helical domain (Kadler et al. 1996), so that may interact with sequences within the triple helix. The type III N propeptide artificially generated in type II procollagen was found to be processed at a considerable rate. The data obtained with the modified construct suggest that the N propeptide of type III
procollagen may interact poorly with the triple-helical domain of type II procollagen, which would explain the enhanced N terminal processing as compared with the wild-type procollagen. A corresponding phenomenon was also observed in the case of type I procollagen when the same modification was made to its proα1(I) and proα2(I) chains (Myllyharju et al. 1997). The cleavage of the type III N propeptide was probably due to non-specific intracellular proteinases, because the proteins studied here were intracellular while the procollagen proteinases are extracellular enzymes (Kielty et al. 1993, Prockop & Kivirikko 1995, Kadler et al. 1996).

The amino acid composition of the purified recombinant type II collagen was very similar to that of the corresponding non-recombinant protein. One major difference between these two proteins was that the hydroxylysine content of the recombinant collagen was only 17 % of that of the non-recombinant protein, and correspondingly the galactosylhydroxylysine and glucosylgalactosylhydroxylysine contents were also low. The hydroxylysine content showed a wide variation between experiments, suggesting that the extent of the interference with the synthesis of endogenous cellular proteins such as lysyl hydroxylase by the baculovirus infection varied considerably between experiments (Jarvis & Summers 1989). In order to increase the hydroxylysine content of the recombinant collagen, insect cells were coinfectd with an additional virus coding for lysyl hydroxylase, which catalyzes the hydroxylation of lysine residues only in the Y positions of the repeating -X-Y-Gly- triplets, and not all -X-Lys-Gly- sequences can be hydroxylated (Kivirikko et al. 1992). This coinfection increased the hydroxylysine content of the recombinant type II collagen to a level that was even slightly higher than that in the non-recombinant protein, the values ranging up to 21 residues/1000 amino acids. It seems likely that the figure of 21 lysines hydroxylated out of the 22 Y-position lysine residues/1000 amino acids (Ala-Kokko et al. 1989) may be the maximal value. The Tm values of the purified low and high hydroxylysine collagens, when determined using CD spectrum analysis, showed no difference, thus indicating that hydroxylysine and its glycosylated forms have no effect on the thermal stability of collagens.

In conclusion, the baculovirus expression system developed here for type II collagen production makes it possible to obtain preparations of this collagen in which the hydroxylysine content ranges from an almost total absence to a maximal value. Such preparations can be used to study the effects of the extent of lysine hydroxylation on fibril formation and other properties of type II collagen.

It has previously been reported that the amount of hydroxylysine varies within the same collagen type in different tissues and even in the same tissue in different physiological and pathological conditions and that the extent of hydroxylation of lysine residues is higher in embryonic tissues than in adult ones (Kivirikko & Myllylä 1982). The amounts of carbohydrate units in collagens also show wide variation between collagen types and within the same collagen type from different tissues (Kivirikko & Myllylä 1979). The precise functions of the hydroxylysine-linked carbohydrate unit are not known, but it has been suggested that they may regulate the lateral packing of collagen molecules within fibrils (Prockop & Kivirikko 1995). Experiments on the de novo generation of type I collagen fibrils from type I procollagen by enzymatic cleavage of the propeptides have demonstrated that a collagen with a higher amount of hydroxylysine and its glycosylated forms makes thinner fibrils than the same protein with
a normal degree of these modifications (Torre-Blanco et al. 1992, Prockop & Kivirikko 1995).

The present data clearly indicate that hydroxylysine and its glycosides have a major influence on type II collagen fibril formation in vitro, as the maximal absorbance of the low hydroxylysine collagen was reached within 5 min, whereas the absorbance of the high hydroxylysine collagen increased for about 600 min and the critical concentration for fibril formation with the high hydroxylysine collagen was about 10 times that with the low hydroxylysine collagen. Electron microscopy of the fibrils also showed major differences, in that those formed by the high hydroxylysine collagen were typically very thin, whereas the low hydroxylysine collagen formed thick fibrils. The highly hydrophilic glucosylgalactose moiety has a length of about 1 nm and is oriented parallel to the backbone of the collagen molecule and shields three or four amino acid residues (Yang et al. 1993). This moiety may inhibit the lateral growth of collagen fibrils because it reduces the surface containing hydrophobic interactions. This suggestion is supported by our morphological findings and also by our data indicating that the final turbidity of the high hydroxylysine collagen was only about one-sixth of that of the low hydroxylysine collagen. Our data also indicate that regulation of the amounts of hydroxylysine and its glycosides may play a major role in regulating collagen fibril formation and the morphology of the fibrils formed in vivo.

6.2 Expression of type I-III procollagens in Pichia pastoris and characterization of the resulting proteins

Yeasts have many advantages as host organisms for the production of heterologous proteins. As a methylotrophic yeast, P. pastoris is capable of metabolizing methanol as its sole carbon source, and it is easy to scale up expression from shaker flasks to high-density fermentor cultures. It was for this reason that the present expression system was chosen for studying large-scale production of the fibril-forming collagen types I-III. The aim was to find out whether the P. pastoris protein expression system could be used to produce hydroxylated collagens, and thus all the genes for the proα chains were introduced into recombinant P. pastoris strains expressing human prolyl 4-hydroxylase subunits.

Most attempts at delineating the biosynthetic pathways of collagens have been made using cultured fibroblasts. Many suggestions have been made about the functions of various proteins in the biosynthesis of procollagens at the levels of transcription, translation and chain assembly. One such protein is Hsp47, a serpin-like chaperone located in the endoplasmic reticulum. Hsp47 binds transiently to collagen types I-V and is believed to be involved in the assembly and packaging of collagens (Nagata 1996). Hsp47 is associated with polysome-bound proα(I) chains (Sauk et al. 1994) and prevents overmodification of type III procollagen in transfected 293 kidney cells (Hosokawa et al. 1998). The present data indicate, however, that Hsp47 is probably not needed for the assembly of triple-helical procollagen molecules, as yeasts have no Hsp47, and as the recombinant collagens expressed here were correctly assembled and folded.
We also wanted to study whether the \textit{P. pastoris} expression system can be used to produce type I procollagen, which consists of two pro\(\alpha_1\)(I) chains and one pro\(\alpha_2\)(I) chain. The results demonstrated that heterotrimeric procollagen molecules can be synthesized and assembled into triple-helical molecules in \textit{Pichia pastoris} cells that express functional prolyl 4-hydroxylase at the same time. The two types of prolyl 4-hydroxylase subunit became assembled into an active \(\alpha_2\beta_2\) tetramer that effectively hydroxylated the two types of pro\(\alpha\) chain of type I procollagen, and the pro\(\alpha_1\)(I) and pro\(\alpha_2\)(I) chains formed heterotrimeric molecules with the correct 2:1 chain ratio. The system reported here using a four-gene expression system may represent the most complex system described so far for the production and processing of any protein in \textit{Pichia pastoris}.

We further studied the \(\alpha\) chains in the pepsin-digested collagen molecules. N-terminal sequencing indicated that the pepsin digestion had removed several residues from the N telopeptides of most of the \(\alpha\) chains but only 1-2 residues from that of the \(\alpha_2\)(I) chain. These N telopeptides are thought to play an important role in fibril formation (Kielty \textit{et al.} 1993, Prockop & Kivirikko 1995, Kadler \textit{et al.} 1996), but a recent study indicates that this role is only seen in the kinetics of the process (Kuznetsova & Leikin 1999), whereas even molecules lacking essentially their entire telopeptides form fibrils that are identical to those formed by the full-length molecules. Fibril formation studies with our recombinant collagen molecules indicated that they all formed native-type fibrils in spite of the lack of a few residues in the termini of their N telopeptides. Our results thus agree with this conclusion.

The N propeptides are believed to have no role in the assembly of the procollagen molecule within the endoplasmic reticulum (Prockop & Kivirikko 1995). We considered it possible that pepsin might digest the N termini of the \(\alpha\) chains less effectively in molecules lacking the N propeptides, because pepsin is an endopeptidase, cleaving proteins at positions within the chain. To study this aspect, we modified the expression constructs for the pro\(\alpha_1\)(I) and pro\(\alpha_2\)(I) chains by removing the sequences encoding the N propeptides. The pC\(\alpha_1\)(I) and pC\(\alpha_2\)(I) chains produced were found to form heterotrimeric pCcollagen molecules with the correct 2:1 chain ratio, and most of the \(\alpha_1\)(I) and \(\alpha_2\)(I) chains indeed had N termini which were identical to those of the corresponding non-recombinant chains. Interestingly, the expression levels of the pCcollagen homotrimers and heterotrimers were found to be 1.5-3 fold higher than the levels of the corresponding procollagen trimers.

Shaker flask culture conditions are not optimal for protein production in \textit{P. pastoris}, due to insufficient \(O_2\), and marked increases in expression levels are usually obtained in fermentors (Romanos \textit{et al.} 1992, Cregg \textit{et al.} 1993). It has previously been reported that the \(K_m\) of \(O_2\) in the prolyl 4-hydroxylase reaction is as high as 40 \(\mu\)M (Kivirikko & Pihlajaniemi 1998), and it has been suggested that the \(O_2\) concentration within the lumen of the endoplasmic reticulum of the yeast cells may be rate-limiting for hydroxylation when a collagen is produced in shaker flask cultures in a \textit{P. pastoris} strain expressing prolyl 4-hydroxylase (Vuorela \textit{et al.} 1997). It has therefore been regarded as possible that the deficiency in the hydroxylation of the recombinant type III collagen seen under such conditions could disappear when the protein is produced in a fermentor equipped with an \(O_2\) supply system (Vuorela \textit{et al.} 1997). The present data verify these suggestions, as the 4-hydroxyproline contents of the recombinant type I-III collagen molecules produced in
the 2-litre fermentor were identical to those of the corresponding non-recombinant proteins and expression of the type I-III procollagens ranged from 200 to 600 mg/l. Many previous studies have demonstrated that the expression levels of various proteins in *P. pastoris* increase markedly with the numbers of DNA copies, at least up to 30-50 copies (Cregg *et al.* 1993, Scorer *et al.* 1994, Werten *et al.* 1999). Furthermore, it has been pointed out that the codon usages of the sequences encoding the human procollagen chains are very different from those encoding various yeast proteins (Werten *et al.* 1999). It would thus seem possible to optimize the current expression levels by using multicopy integrants and DNA sequences with optimized codon usages for the production of very large amounts of recombinant human type I collagen for various medical applications.
7 References


Prockop DJ, Sieron AL & Li SW (1998) Procollagen N-proteinase and procollagen C-proteinase. Two unusual metalloproteinases that are essential for procollagen processing probably have important roles in development and cell signaling. Matrix Biol 16: 399-408.


Yao Y, Zhou Y & Wang C (1997) Both the isomerase and chaperone activities of protein disulfide isomerase are required for the reactivation of reduced and denatured acidic phospholipase A2. EMBO J 16: 651-658.