Isolation and Characterisation of Four Cathepsin B-like Cysteine Protease Genes from the Free Living Nematode *Caenorhabditis elegans*

Christopher Geoffrey Carson Larminie

Ph.D.
University of Glasgow
Wellcome Unit of Molecular Parasitology
December 1995

Contents

Chapter 1
General Introduction

Chapter 2
Materials and Methods

Chapter 3
Cloning Four Cathepsin B-Like Genes From *C.elegans*

Chapter 4
Characterisation of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

Chapter 5
The Temporal and Spatial Expression Patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

Chapter 6
Concluding Discussion

Abstract

The cathepsin B cysteine protease enzyme performs a role in protein turnover and degradation within the lysosomes of vertebrates. This proteolytic enzyme is also thought to perform related roles in the processing of antigens and protein precursors, as well as a role in bone resorption. A pathological role for cathepsin B in tumour cell invasion has also been suggested. Enzymes with cathepsin B-like activities are thought to be excreted and/or secreted by a variety of parasitic nematode and trematode species. To date, multigene families with the potential to encode cathepsin B-like enzymes have only been reported in parasitic nematode and trematode species, suggesting that these enzymes may be important for parasitism by these species. The work presented in this thesis demonstrates that the free living nematode species, *Caenorhabditis elegans*, also possesses a cathepsin B-like multigene family indicating that such multigene families are not unique to parasitic nematode and trematode species. Four genes with homology to vertebrate cathepsin B were isolated from the genome of *C.elegans* and were named *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. Phylogenetic analysis clusters the proteins encoded by these four genes with known cathepsin B enzymes and away from other, related enzymes such as cathepsins H and L. Since the four genes possess distinct genomic architectures, they appear to have arisen from ancient gene duplication events. This is supported by phylogenetic analysis which clusters the predicted proteins encoded by these four genes and *cpr-1*, a previously isolated *C.elegans* cathepsin B-like gene (Ray and McKerrow, 1991), into three groups which are almost as diverged from one another as each is to the vertebrate cathepsin B enzymes. The expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as *lacZ* reporter transgene fusions in transgenic worms suggests that these four genes are all exclusively expressed in the intestinal cells of *C.elegans*. Analysis of the temporal expression patterns of these four genes using semi-quantitative reverse transcription polymerase chain reaction indicates that the four genes exhibit distinct but overlapping temporal patterns of expression during *C.elegans* development. These results suggest a differential requirement for cathepsin B-like enzymes, or combinations of enzymes, within the intestine during *C.elegans* development.

Acknowledgements

I would like to thank Dr. Iain Johnstone, my supervisor, for his encouragement and support over the last three years. I would also like to thank the other people involved in *C.elegans* research at Glasgow University and in particular Dr. John 'Boy' Gilleard. I am also grateful to the people of the Wellcome Unit of Molecular Parasitology with whom I have enjoyed working. I would like to acknowledge the financial support of the Scottish Office Agriculture and Fisheries Department. I am especially grateful to A.J.M. for all her support over the last three years, particularly during the writing of this thesis.

Declaration

This thesis and the results presented in it are entirely my own work, except where indicated.

Abbreviations

bp	base pair
BSA	bovine serum albumen
cDNA	complementary deoxyribonucleic acid
cm	centimetre
cpm	counts per minute
DNA	deoxyribonucleic acid
E/S	excretory/secretory
EST	expressed sequence tag
IVR	*in vitro* released
kb	kilobase
kcal	kilocalories
kDa	kiloDalton
l	litre
lbs	pounds
M	Molar
Mb	megabase
MCS	multiple cloning site
min	minute
ml	millilitre
mM	Millimolar
mm	millimetre
µm	micrometre
µl	microlitre
µg	microgram
ng	nanogram
OD	optical density
PCR	polymerase chain reaction
pfu	plaque forming unit
pg	picogram
RACE	rapid amplification of cDNA ends
rDNA	ribosomal deoxyribonucleic acid
RNA	ribonucleic acid
rpm	revolutions per minute
RT-PCR	reverse transcription polymerase chain reaction
s-q rtPCR	semi-quantitative reverse transcription polymerase chain Reaction
ssDNA	single stranded Deoxyribonucleic acid
UTR	untranslated region
UV	ultraviolet
V	volt
YAC	yeast artificial chromosome
X-gal	5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside

Chapter 1

Chapter 1

General Introduction

1.1. The Phylum Nematoda

The phylum Nematoda is large, containing 10,000 described nematode species, and its members are among some of the most widespread and numerous of all multicellular animals. Members of this phylum extend from the polar regions to the tropics and inhabit all types of marine and terrestrial environments including; high altitudes, hot springs, deserts and oceans. Parasitic nematodes are equally diverse, displaying all the different degrees of parasitism and attacking virtually all groups of plants and animals. Despite the diverse environments and very different lifestyles of different nematode species, all are morphologically and anatomically very similar. All nematode species have a long slender body and a circular cross section (hence the name 'roundworms'). The general body plan comprises two concentric tubes separated by the pseudocoelomic space. The outer tube consists of the cuticle, the hypodermis, muscle and nerve cells while the inner tube consists of the intestine and its lumen. The gonads of adult worms are contained in the pseudocoelomic space. For a more in-depth overview of the characteristics of the phylum Nematoda, see Barnes (Barnes, 1980).

1.2. *Caenorhabditis elegans* biology

The free-living soil nematode, *Caenorhabditis elegans*, has been the subject of extensive studies and is the most well characterised of all nematode species. It is a filter feeder whose diet consists primarily of bacteria. *C.elegans* can be easily maintained in the laboratory on agar plates or in liquid culture using *Escherichia coli* as a food source (Brenner, 1974) and the life cycle takes only approximately 3½ days to complete under optimal conditions. The two sexes, hermaphrodite and male, are a similar size, approximately 1mm in length and 70µ in diameter. However, the two sexes can be distinguished by virtue of different anatomical features (see Section 1.2.2). *C.elegans* propagates primarily by hermaphroditic self-fertilisation since hermaphrodites cannot fertilise one another and males, which are capable of fertilising hermaphrodites, occur in the population only at a low frequency (approximately 1/700).

1.2.1. *C.elegans* development

The life cycle of *C.elegans*, like all other nematode species, involves four larval stages, named L_1 - L_4. Eggs are fertilised internally and early embryogenesis occurs within the parent. The later stages of embryogenesis occur after the eggs have been laid. Embryogenesis is rapid, taking only approximately 14 hours to complete in optimal conditions. After hatch, the morphology of the resulting 1st stage larvae (L_1) is superficially quite similar to the adult. Three further larval stages occur (L_2, L_3 and L_4) prior to generation of the adult approximately three days after hatch. The end of each larval stage is punctuated by a moult. At each moult, the old cuticle is shed to reveal a newly synthesised cuticle beneath. During shedding of the old cuticle, pharyngeal pumping stops for a brief period known as lethargus. Under conditions of overcrowding or starvation, the L_2 larvae can enter an alternative developmental pathway to generate dauer larvae which are long-lived and resistant to unfavourable conditions (Riddle, 1988). In more favourable conditions, these dauer larvae moult to L_4 stage larvae and resume normal development.

The body wall of *C.elegans* and the shell of eggs and developing embryos are transparent. This feature, in conjunction with Nomarski microscopy (a high resolution, non-destructive light microscopy technique), has made it possible to describe the complete embryonic (Sulston *et al.,* 1983) and post-embryonic (Sulston and Horvitz, 1977) cell lineages of *C.elegans*. These studies followed individual nuclei in live animals and compared the resulting observations to the cellular anatomy of the worm, determined at both the light and electron microscope levels. These studies have revealed that both the embryonic and post-embryonic cell lineages are highly invariant, with each lineage rigidly determined and giving rise to a fixed number of cells of strictly specified fates. Although the relationship between cell ancestry and cell fate is fixed, there is little correlation between them. Thus, the six founder cells, AB, MS, E, C, D and P_4, generated by the first four cleavages, do not correspond exactly to specific germ layers. However, three founder cells do give rise to 'pure' clones; E gives rise exclusively to the intestine, D produces body wall muscle (but not all body wall muscle) and P_4 gives rise to the germ-line.

As a result of the cell lineage studies described above, the entire period of *C.elegans* development is very well characterised. Embryogenesis can be divided into two stages of approximately the same length; first, the cell proliferation and organogenesis stage and second, the morphogenesis stage. During the first stage, cell divisions, cell movements and some cell deaths take place. By the end of this phase the gastrulated embryo is spheroid with a fixed number of cells (approximately 550) whose fates are rigidly determined. During the second phase, cell proliferation stops almost completely and morphogenesis occurs. During this stage, the embryo elongates more than threefold, neural processes grow out and interconnect, the embryo begins to move actively and cuticle synthesis occurs.

The L_1 stage larvae comprise approximately 550 somatic nuclei in both hermaphrodites and males. Both sexes possess a gonadal primordium at this larval stage which comprises four cells. The number of somatic nuclei increases to 959 in the mature adult hermaphrodite and 1031 in the mature adult male. These additional nuclei are generated both by division of non-gonadal somatic blast cells and by proliferation of the gonadal primordium during post-embryonic development.

The non-gonadal somatic blast cells make up approximately 10% of the non-gonadal nuclei of both hermaphrodite and male L_1 stage larvae (Sulston and Horvitz, 1977). As a result of division of these blast cells, the number of non-gonadal somatic nuclei increases to about 810 in the mature adult hermaphrodite and 970 in the male. In both sexes, the daughter cells arising from these divisions contribute to the hypodermis, nervous system, intestine and musculature. Most of the additional male-specific cells are located in the specialised structures of the male tail which are required for copulation.

The remaining somatic nuclei of the mature adult hermaphrodite and mature adult male are generated by proliferation of the gonad. The hermaphrodite and male gonad develops during the larval stages, from a gonadal primordium comprising four cells in the L_1 stage. Of these four morphologically similar cells, two contribute to the somatic gonad and two to the germ-line tissue. The somatic gonad of both sexes develops according to an invariant pattern of cell divisions (Kimble and Ward, 1988), resulting in approximately 140 nuclei contributing to the hermaphrodite gonad and 60 nuclei to the male gonad. Though gonadogenesis occurs throughout larval development, the structures of the gonad only begin to be easily visible around the late L_3 stage. Sexual maturation continues to the end of the L_4 stage, culminating in the opening of the hermaphrodite vulva and morphogenesis of the male tail at the final moult.

1.2.2. *C.elegans* anatomy

The complete anatomy of *C.elegans* has been determined at the cellular level mostly using reconstructions of electron micrographs of serial sections. This information, in conjunction with the invariance of cell number and cell fate (discussed above), has allowed every cell in the worm to be identified and given a unique label. The anatomy of *C.elegans* is typical of all nematode species and essentially consists of two concentric tubes separated by pseudocoelomic space. The outer tube is made up of a cuticle, hypodermis, neurones and muscle which all surround the pseudocoelom. The inner tube consist of the intestine and its lumen. The gonads of both sexes are contained within the pseudocoelomic space (Figure 1.1A).

The cuticle is made up of collagen, organised into three main layers, and is synthesised by an underlying external epithelium called the hypodermis. Accordingly, the hypodermis extends over the entire surface of the worm, including parts of the lumen of the pharynx and anus. Two elevated longitudinal ridges, called alae, mark the lateral surfaces of the adult cuticle. These alae are synthesised by seam cells, a subset of hypodermal cells which underlie the alae. On solid surfaces, *C.elegans* crawls on one side with the alae contacting the surface of the medium. Movement itself is attained using four strips of striated body wall muscle running along the length of the animal (Figure 1.1A).

The alimentary canal comprises the pharynx, the intestine and the rectum. The pharynx functions to ingest, concentrate and process food prior to pumping it into the intestine. Accordingly, the pharynx is made up of muscles, epithelial cells and nerves. The lumen of the pharynx is lined with cuticle and this cuticle is continuous with the body cuticle. The pharyngeal cuticle possesses several specialised structures the most obvious of which are the knob like structures in the terminal bulb (Figure 1.1B), which grind up food. The pharynx is connected to the intestine via the pharyngeal-intestinal valve. The intestine is comprised of two rows of eight cells (int2-int9) and an anterior ring of four cells (int1). The four cells comprising int1 represent the point of attachment of the intestine to the cells of the pharyngeal-intestinal valve. The intestine is attached to the rectum via the intestinal-rectal valve, a very similar structure to the pharyngeal-intestinal valve. The rectum itself is made up of three pairs of endothelial cells. The rectum is associated with three sets of muscles required for excretion which are coupled by gap junctions and innervated by a single neuron.

The hermaphrodite and male gonads have very different structures (Figure 1.1B). The hermaphrodite gonad has two lobes, one extending anteriorly and the other posteriorly from the centre of the worm. Each lobe is U-shaped and comprises an ovary (at the distal end), an oviduct and a spermatheca (at the proximal end). At the distal end, the ovaries are syncytial and contain germ-line nuclei. As the nuclei move proximally, they progress through the stages of meiosis and are in diakinesis by the time they reach the oviduct. By this stage, each of the germ-line nuclei are enclosed by membranes to form large oocytes. The oviduct terminates at the spermatheca, containing amoeboid sperm. The two spermathecae (one from each lobe) are connected by a common uterus which contains fertilised eggs. The uterus opens to the outside via a vulva protruding from the ventral surface of the adult hermaphrodite.

The male gonad is a single lobed, U-shaped structure which extends anteriorly, flips back on itself and then extends posteriorly to the cloaca. At the end distal to the cloaca, the germ-line nuclei are mitotic. These nuclei become meiotic, and advance through the various stages of meiosis, as they travel proximally to the seminal vesicle. The sperm are stored in the seminal vesicle until copulation, when they are released via the vas deferens and cloaca. The male tail has specialised neurons, muscles and hypodermal structures required for mating, including two spicules that are inserted into the hermaphrodite vulva to aid the transfer of sperm.

The nervous system is the most complex organ of *C.elegans*, accounting for 37% and 46% of somatic nuclei in hermaprodites and males, respectively (Chalfie and White, 1988). The nervous system, including all the neural connectivities, has been completely reconstructed from electron micrographs of serial sections (White *et al.,* 1986). The system is divided into two almost independent units, the 20 nerve cells of the pharynx and the nerve cells of the rest of the body. Most of the cell bodies of these neurons are organised into ganglia. Most of the neuronal processes from these cells form a ring around the outer surface of the pharynx (the nerve ring) or contribute to nerve bundles running the length of the body (most notably, the dorsal and ventral nerve cords). Most of the sensory neurons run anteriorly from the nerve ring to sensory structures in the head (sensilla) while most motor neurons run posteriorly from the nerve ring to the body wall muscle. A more detailed overview of *C.elegans* anatomy is given by J.White (1988).

1.2.3. *C.elegans* genetics

The haploid genome of *C.elegans* is relatively small, with approximately 10^8 bp (100 megabases) of DNA. Wild-type hermaphrodites possess five pairs of autosomes (A) and a pair of X chromosomes (XX) while males possess five pairs of autosomes and only a single X chromosome (XO). The ratio of X chromosomes to autosomes has been shown to be the primary sex-determining signal of *C.elegans* by altering the X:A ratio (Hodgkin *et al.*, 1979; Madl and Herman, 1979).

C.elegans has been the focus of much classical genetic analysis and its mode of reproduction has been very useful in this respect. On one hand, genetic crosses, which are essential for classical genetic analysis, can be performed between hermaphrodites and males. On the other hand, the predominant self-fertilising hermaphroditic mode of reproduction allows easier isolation of recessive mutations, since homozygous worms appear in the F2 generation automatically, without the need for sibling crosses. Large scale genetic screens have been performed using chemical mutagenesis and radiation. Indeed, one of the earlier screens (Brenner, 1974) identified at least 77 genetic loci, distributed over the 6 linkage groups. Since then, over 1200 genetic loci have been identified using classical genetic approaches (S.Martinelli pers. comm.).

1.3. *C.elegans* as an experimental system

The features of *C.elegans* biology discussed above make it a very useful system for experimental analysis. The worm is easily maintained in the laboratory and has a short life cycle, facilitating genetic analysis. The worm is sufficiently simple to allow thorough analysis, as evidenced by the work performed on the cell lineage and anatomy of this worm, yet is sufficiently complex to allow studies of cell interaction, animal development and behaviour. Furthermore, extensive classical genetic analysis of *C.elegans* has contributed substantially to our understanding of how individual genes participate in the development and behaviour of this nematode species.

There are several additional features which make *C.elegans* an extremely powerful experimental system. First, methods for transforming *C.elegans* have been developed which allow genetically engineered genes to be reintroduced (Stinchcomb *et al.*, 1985; Fire, 1986). Second, a physical map has been generated which covers over 95% of the *C.elegans* genome and this map has been aligned with the genetic map. The alignment of the physical and genetic maps, in conjunction with *C.elegans* transformation techniques, has greatly facilitated the isolation and cloning of mutationally defined genes using mutant rescue approaches. The physical map and its uses are discussed in more detail in Chapter 3, Sections 3.1.1-3.1.3. Third, the ongoing *C.elegans* genome sequencing project (discussed in Chapter 3, Section 3.1.2) will elucidate the structure of the entire *C.elegans* genome and will allow identification of those genes not defined by mutations. Fourth, a powerful technique for reverse genetics is available for *C.elegans* and will allow the functions of many of the genes identified by the genome sequencing project to be determined. The technique uses polymerase chain reaction (PCR) to detect insertions of the Tc1 transposon in, or near, a gene of interest. In brief, the method requires the use of mutator strains of *C.elegans* which activate Tc1 transposition and excision at high frequency in the germ-line. Genomic DNA from a number of populations of these worms are screened by PCR using primers specific to the target gene and to Tc1. Because genomic DNA is used in the PCR screen, it is necessary to use sibling selection, where siblings of the analysed animals are kept alive. Once a population of worms containing the desired Tc1 insertion has been identified, the siblings are divided into smaller sub-populations and allowed to progress through 1 generation before being analysed. By repeating the PCR screen and sibling selection steps a few times, it is possible to identify single worms containing the desired Tc1 insertion (Rushforth *et al.*, 1993). Furthermore, the ability to retrieve up to 50% viable animals from cultures of *C.elegans* stored at -70°C has permitted the construction of frozen Tc1 insertion mutant banks (Zwaal *et al.*, 1993). Such Tc1 insertion mutant banks allow the rapid isolation of Tc1 insertion alleles in almost any gene. For example one such bank has already isolated Tc1 insertion alleles for over 70 different genes (H.C.Korswagen pers. comm.).

Antisense RNA techniques can also be used to obtain mutant phenotypes for genes identified by DNA sequencing. In *C.elegans*, inhibition of gene expression by antisense RNA has been achieved simply by injecting antisense RNA into the hermaphrodite gonad. Such experiments have demonstrated that this approach is capable of phenocopying known null mutant phenotypes of the genes being analysed. For example, antisense RNA techniques have been used to phenocopy mutations in the *par-1* maternal effect gene. These experiments demonstrated that embryos from wild-type *C.elegans* injected with *par-1* antisense RNA arrested development with a *par-1* terminal phenotype, with no morphogenesis, no intestinal cells and excess of pharynx (Guo and Kemphues, 1995). This approach has also been used to study zygotically expressed genes that act later in development. For example, antisense RNA to the CeMyoD gene (which encodes a myogenic basic helix-loop-helix protein) phenocopied a null mutation of this gene, generating embryos with a dumpy phenotype which did not develop further than the twofold stage (M.Park and M.Krause pers. comm.). Specific antisense RNA mediated gene inhibition has also been obtained by injecting *C.elegans* with constructs containing target DNA in the reverse orientation under the control of a heterologous promotor (Fire *et al.*, 1991).

The extensive knowledge of *C.elegans* biology, together with the experimental tractability of this nematode species, make it a powerful tool for isolating genes and determining gene function, using both classical and reverse genetic approaches.

1.4. *C.elegans* as a model organism

The simplicity and experimental tractability of *C.elegans*, discussed above, has facilitated the accumulation of a substantial amount of information regarding its biology. Much of this information has revealed that *C.elegans* biology shares much more in common with higher eukaryotes than might be expected from its simple body plan, supporting the contention that information obtained from this nematode species will be relevant to understanding mammalian biology. This is demonstrated by the numerous genes isolated from *C.elegans* and vertebrates which show extensive homology to one another, suggesting that *C.elegans* may share many cellular pathways in common with higher eukaryotes. Two examples illustrate this well. First, several genes necessary for *C.elegans* vulval differentiation encode proteins which are homologous to mammalian signal transduction proteins. The *lin-3* gene encodes an epidermal growth factor (EGF)-like molecule that is an inductive signal from the anchor cell of the gonad and which induces three ectodermal precursor cells to adopt a vulval fate (Hill and Sternberg, 1992). The *let-23* gene encodes a receptor tyrosine kinase of the EGF receptor subfamily which is thought to be the receptor for the *lin-3* product (Aroian *et al.,* 1990). The *let-60* gene encodes a *ras* protein that acts downstream of *let-23* to control vulval fate (Han and Sternberg, 1990; Beitel *et al.,* 1990). The *lin-45* gene encodes a homologue of the mammalian *raf* family of serine/threonine kinases (Han *et al.,* 1993). The order of action of these genes in vulval development has been studied genetically and suggests that *let-23* positively regulates *let-60* which in turn positively regulates *lin-45*. Thus, these results suggest that *C.elegans* utilises a similar receptor tyrosine kinase signalling pathway for regulating vulval development as mammals do for regulating cellular responses to growth-factors.

In the second example, two genes involved in the *C.elegans* programmed cell death pathway have been shown to be functional homologues of mammalian cell death genes. The *C.elegans ced-3* gene has sequence similarities to the interleukin-1ß-converting enzyme (ICE). Over expression of either the murine ICE gene or *ced-3* gene in Rat-1 cells causes these cells to undergo programmed cell death (Miura *et al.,* 1993). The *C.elegans ced-9* gene is required to protect cells from programmed cell death. Accordingly, loss of function mutations of this gene result in cells which should normally live, to enter the programmed cell death pathway. *ced-9* encodes a protein with sequence similarities to the mammalian *bcl-2* proto-oncogene. Over expression of *bcl-2* in *C.elegans* can mimic the protective effect of *ced-9* on *C.elegans* cell death and can prevent the ectopic cell deaths that occur with *ced-9* loss of function mutants (Hengartner and Horvitz, 1994). Such examples suggest that many cellular pathways may be conserved between *C.elegans* and mammals and therefore support the use of *C.elegans* as a general model for metazoan biology.

1.5. Cathepsin B

For several reasons, we decided to study the cathepsin B-like genes of *C.elegans*. Before discussing these reasons, I will give some background information on the current state of knowledge of cathepsin B activity and function. Cathepsin B is a proteolytic enzyme associated with the lysosomes. All proteolytic enzymes perform the same biochemical function of hydrolysing the peptide bond between two amino acid residues. Some proteases possess endopeptidase activity, cleaving peptide bonds towards the centre of a polypeptide chain, while others act as exopeptidases, cleaving bonds near the termini of a polypeptide chain. These activities are not mutually exclusive. Therefore, for the purposes of classification, all enzymes capable of acting as endopeptidases are classed as endopeptidases irrespective of whether or not they can also act as exopeptidases.

Endoproteolytic enzymes (endopeptidases) cannot be classified on the basis of substrate specificity because these enzymes do not recognise a specific substrate but rather have preferences for certain amino acid sequences in certain conformations, giving rise to a number of target substrates. Accordingly, endoproteolytic enzymes are classified according to their mechanism of catalysis. The International Union of Biochemistry currently recognises four mechanistic classes of proteolytic enzymes; the cysteine, serine, aspartic acid and metallo proteases. Accordingly, cathepsin B has been classified as a member of the cysteine protease class of endoproteolytic enzymes.

1.5.1. Papain

Papain, from *Carica papaya*, was the first recognised member of the cysteine protease class of endopeptidases. The information gained from biochemical analysis of papain was essential for the understanding of cathepsin B and therefore I will give a brief overview of the pertinent discoveries made from the studies of papain. Papain requires a free sulphydryl (-SH) group for activity and maximal activity is only obtained in the presence of thiol compounds, such as glutathione, or reducing agents, hence the terms 'thiol-dependent' or 'cysteine' are used to describe this protease, and other proteases of this class.

The amino acid sequence of the single polypeptide chain of papain was reconstructed from amino acid sequencing of peptides generated from tryptic, chymotryptic and peptic digests (Light *et al.*, 1964). During these experiments, the use of ^{14}C-labelled iodoacetic acid, which covalently modifies the active -SH group, demonstrated that the cysteine residue at position 25 possesses the active thiol group of papain. The tertiary structure of crystallised papain was determined by X-ray diffraction (Drenth *et al.*, 1968) and mostly agreed with the predicted primary amino acid sequence of papain. The X-ray data for papain indicates that this single 212 amino acid peptide is folded into an ellipsoid particle and possesses a binuclear structure, generated by two hydrophobic cores. The active site, and the region of interaction of the enzyme with its substrate, occurs on the surface of the papain molecule in a large groove, situated between the two hydrophobic cores of the molecule. The sulphydryl side chain of the active site cysteine residue (position 25) is found in the groove closely associated with the imidazole ring of a histidine residue (position 159). The imidazole ring is in turn hydrogen bonded to the side chain of an asparagine residue (position 175), preventing any movement of this ring. The currently accepted mechanism of catalysis by papain, and all other cysteine proteases, is shown in Figure 1.2. This mechanism is based on the X-ray structure and a substantial amount of chemical and kinetic studies performed with papain. A more detailed overview of the structure and biochemical characterisation of papain is given by Drenth *et al* (1971) and by Glazer and Smith (1971).

Papain possesses a large catalytic site of about 25 Å which can be divided into seven ‘subsites’, each accommodating one amino acid residue of the substrate. These ‘subsites’ are positioned either side of the active site, four on one side and three on the other (Schechter and Berger, 1967). A schematic diagram of the papain active site, showing the scheme for numbering the subsites is shown in Figure 1.3. These subsites are an important basis for the substrate specificity of papain and other related cysteine protease enzymes, including cathepsin B. For example, the S_2 subsite of papain specifically interacts with a phenylalanine (Phe) side chain (Schechter and Berger, 1968) which results in peptides containing a Phe residue in the appropriate position having increased susceptibility to hydrolysis by papain.

As a result of the large catalytic site of papain and other related cysteine proteases, small synthetic substrates cannot accurately reflect the protein substrate specificity of these enzymes because these synthetic substrates will only interact with a few of the subsites. Furthermore, even the use of larger peptides which interact with all the subsites will not accurately reflect the protein substrate specificity of these enzymes because recognition of such target sequences within the protein substrate will be dependent on the conformation of the protein. Thus, a protein substrate may possess several peptide sequences recognised by papain but only some, or none, will be susceptible to cleavage by this enzyme, resulting in what is called 'limited proteolysis'.

1.5.2. Isolation and biochemical characterisation of cathepsin B

The cathepsin B enzyme was first detected from bovine spleen as an activity capable of hydrolysing the synthetic substrate benzoylarginine amide in the presence of cysteine (Greenbaum and Fruton, 1957). This activity was later purified from rat liver lysosomes by chromatography and was shown to correspond to a 25 kDa protein which was named cathepsin B (Otto, 1971). Initial characterisation of cathepsin B revealed that this enzyme, and other lysosomal cysteine proteases including cathepsins L and H, show optimal activities at slightly acidic pH and are irreversibly inactivated by alkaline conditions (Barrett and Kirschke, 1981). Extensive biochemical characterisation of cathepsin B isolated from a number of vertebrates, and determination of the three dimensional structure of human liver cathepsin B, revealed that this enzyme is a cysteine protease with substantial homology to papain. Some aspects of cathepsin B revealed by these studies are discussed in the following paragraphs.

Substrate and inhibitor studies have been used extensively to characterise the activity of cathepsin B. For example, the use of synthetic substrates has demonstrated that the S_2 subsite of cathepsin B has a slight preference for basic residues or phenylalanine and the S_1 subsite disfavours bulky side chains (Shaw *et al.,* 1983). Though information obtained from using small synthetic substrates cannot be used to infer the protein substrates of cathepsin B, it can be exploited in the design of substrates and inhibitors that are specific to cathepsin B. For example, the ability of cathepsin B to cleave synthetic substrates containing a pair of basic amino acid residues (Barrett and Kirschke, 1981) has proved to be a characteristic feature of this enzyme, and a variety of fluorogenic synthetic substrates which contain two arginine residues, such as Z-Arg-Arg-

NHMec (benzyloxycarbonyl-Arg-Arg-4-methylcoumarin-7-ylamide), have been designed to take advantage of this (Barrett and Kirschke, 1981). Such synthetic substrates are very useful for discerning between different cysteine protease activities, or determining the relative activity of different cysteine proteases within a single sample.

Protein substrates have also been used in the characterisation of cathepsin B and have revealed some interesting features of cathepsin B activity. Most importantly the use of glucagon (Aronson and Barrett, 1978) and rat muscle aldolase (Bond and Barrett, 1980) as substrates has revealed that cathepsin B appears to be unique among the cysteine proteases in its ability to act both as an endopeptidase and as an exopeptidase (peptidyldipeptidase).

Inhibitor studies have been widely used in the classification of endopeptidases because they are capable of revealing the chemical nature of the catalytic groups in a way that substrates cannot. Cathepsin B activity has been shown to be inhibited by a number of cysteine protease-specific inhibitors, indicating that it is a member of the cysteine protease class of endopeptidases. For example the cysteine protease inhibitor E-64, and derivatives of this, have been used to study inhibition of cathepsins B, H and L (Barrett *et al.,* 1982). A large number of synthetic inhibitors have been developed which are specific for particular enzymes. These inhibitors generally comprise one or two amino acid residues linked to a reactive group. The amino acid residues of these inhibitors bind to the subsites comprising the active site and the reactive group inactivates the active thiol group of the enzyme. Thus, it is possible to obtain inhibitors which preferentially inactivate different cysteine proteases by screening those enzymes with inhibitors that possess different amino acid combinations (Kirschke and Barrett, 1987). By their nature, such inhibitors also provide information on the enzyme's preference for certain amino acid residues and therefore provide valuable information regarding the specificity of the subsites comprising the active site of the enzyme. One such group are the peptidyl diazomethanes, some of which show very large differences in the inactivation rates for cathepsin B and cathepsin L (Shaw *et al.,* 1983). Inhibitors specific for a particular enzyme are valuable experimental tools since they can be used to study the biological effects of inhibiting that enzyme. This approach has been used extensively for determining the biological roles of cathepsin B and related enzymes and will be discussed later.

Amino acid sequencing has elucidated the complete primary amino acid sequences of rat liver cathepsin B (Takio *et al.*, 1983) and human liver cathepsin B (Ritonja *et al.*, 1985) and the partial amino acid sequences of porcine liver cathepsin B (Takahashi *et al.*, 1979; Takahashi *et al.*, 1980) and bovine spleen cathepsin B (Pohl *et al.*, 1982). A comparison of the primary amino acid sequences of rat and human liver cathepsin B reveals a very high level of similarity (Ritonja *et al.*, 1985). Both enzymes are 252 amino acids long and share 83.7% identity, suggesting that they represent functional homologues. The data obtained from amino acid sequencing has demonstrated that crystallised cathepsin B from pig liver, beef spleen and rat liver is present in a mixture of single and two chain forms while human liver cathepsin B is almost exclusively in the two chain form. The two chain form is generated by limited proteolysis of the single chain form and occurs at the equivalent position for all these cathepsin B enzymes, between an asparagine and valine residue, to generate a light chain and heavy chain. These two chains are linked covalently by disulphide bonds. It is unclear whether this limited proteolysis is part of the normal processing of cathepsin B, a step in the degradation of cathepsin B, or an artefact of the purification method.

The three-dimensional structure of crystallised human liver cathepsin B has been determined (Musil *et al.*, 1991). Comparison of the three dimensional structures of human liver cathepsin B and papain reveal that they are very similar to one another in their overall structure, including the catalytic site and active site cleft. However, the authors suggested that there were some differences which might explain the different activities of these two enzymes. Most notably, cathepsin B possesses a novel loop containing two adjacent histidine residues (His-110 and His-111) which is inserted into the active site. This loop may explain the C-terminal peptidyldipeptidase (exopeptidase) activity of cathepsin B since it is thought to favour binding of peptide substrates with two residues carboxy-terminal to the bond which is cleaved. In such circumstances, the adjacent histidine residues of the loop are thought to anchor the carboxy-terminal of the peptide substrate. A second difference observed is the presence of a glutamic acid residue in the S_2 subsite (Glu-245) which would explain the characteristic preference of cathepsin B for basic residues at the P_2 site of peptide substrates. Accordingly, mutation of the His-111 residue of recombinant rat cathepsin B to glutamine results in much reduced exopeptidase activity and increased endopeptidase activity with respect to wild-type rat cathepsin B while mutation of Glu-245 to glutamine or alanine alters the substrate specificity of recombinant rat cathepsin B (Hasnian *et al.,* 1992). This is a clear example of how structure relates to function. Accordingly, information obtained from the three-dimensional structure of human liver cathepsin B may be useful not only for predicting the structure of the predicted products of cloned cathepsin B-like genes for which no biochemical data is as yet available, but also for identifying residues likely to alter the activity and specificity of the predicted proteins.

1.5.3. Processing of cathepsin B

Comparison of the predicted amino acid sequence of human and rat cathepsin B (obtained by DNA sequencing of cDNA clones encoding these enzymes) to the primary amino acid sequence of mature human and rat cathepsin B (determined by amino acid sequencing) has revealed that these enzymes are translated as precursors (Chan *et al.,* 1986; San Segundo *et al.,* 1985). This comparison revealed that at least two cleavages must occur to release an 81 amino acid N-terminal proregion and a C-terminal hexapeptide. The predicted precursor of mouse cathepsin B, determined by sequencing of its cDNA clones (Chan *et al.,* 1986), is very similar to those of human and rat cathepsin B, suggesting that similar processing events occur for this enzyme. Furthermore, the precursors of these three cathepsin B enzymes all possess a seventeen residue predominently hydrophobic domain at the amino-terminus. Such signal sequences (known as prepeptides) are required to sequester the nascent protein within the lumen of the endoplasmic reticulum and are usually rapidly removed after synthesis. Thus at least two cleavage events are thought to be required to remove first the prepeptide and then the proregion from the amino terminus of these enzymes.

Recent studies using recombinant human cathepsin B expressed in yeast, suggest that procathepsin B is activated by an autocatalytic mechanism. For example, the proteolytic processing of a non-activatable form of recombinant human cathepsin B (generated by site-directed mutagenesis of the active site cysteine residue) by lysosomal enzymes in microsomal fractions is inhibited by the cathepsin B-selective CA-074 inhibitor (Mach *et al.,* 1993), suggesting that cathepsin B activity is required for procathepsin B processing. Furthermore, activation and proteolytic maturation of human recombinant cathepsin B has been shown to occur primarily by a concentration-independent process, indicating an intramolecular mechanism whereby procathepsin B is capable of processing itself (Mach *et al.,* 1994a). In these experiments, a single cleavage occurred during the processing of procathepsin B to yield an intact proregion. This may be physiologically relevant since the proregion is thought to be able to act as a potent reversible inhibitor and may stabilise mature cathepsin B *in vivo* (Mach *et al.,* 1994b).

These results are at odds with previous studies which suggest that cathepsin B is activated by other enzymes. For example, pepstatin a potent inhibitor of cathepsin D inhibits activation of both cathepsin B and L (Nishimura *et al.,* 1988). Other studies, using inhibitors for each of the four classes of endoprotease, suggested that metalloprotease activity is required for processing of cathepsin B and L in macrophages (Hara *et al.,* 1988). Such anomalies may simply reflect a more complex situation *in vivo,* where cathepsin B may be capable of both auto-activation and activation by other classes of enzyme. Furthermore, whilst some activation pathways may leave the proregion intact and allow it to act as an inhibitor others may degrade it completely. Thus, the type of activation pathway used may affect subsequent activity of cathepsin B.

Studies with mammalian cells have demonstrated that glycosylation is an integral part of lysosomal trafficking, with mannose-6-phosphate being the recognition signal for transport of soluble proteins to the lysosomes (see Chapter 4, Section 4.3.3.2). Accordingly, cathepsin B from human liver, rat liver and porcine spleen are all glycosylated at asparagine-111 (Ritonja *et al.,* 1985; Takio *et al.,* 1983; Takahashi *et al.,* 1984). However, complete hydrolysis of human liver cathepsin B suggests that it possesses only a small carbohydrate prosthetic group compared to rat liver and porcine spleen cathepsin B (Barrett, 1977; Taniguchi *et al.,* 1985; Takahashi *et al.,* 1984). Furthermore, three different carbohydrate structures were identified for rat liver cathepsin B (Taniguchi *et al.,* 1985) and two different structures for porcine cathepsin B (Takahashi *et al.,* 1984). Whilst these two enzymes have one type of carbohydrate structure in common, none of these structures are like the high-mannose type structures normally linked to lysosomal enzymes. Together, these results suggest that vertebrates may possess different isoenzymes of cathepsin B which possess different carbohydrate moieties. Given the role of mannose-6-phosphate as a lysosomal targeting signal, these differences may result in the targeting of different cathepsin B isoenzymes to different subcellular compartments.

1.6. Tissue distribution of cathepsin B

Cathepsin B enzymes have been isolated from human liver, rat liver, porcine spleen, bovine spleen, rabbit liver and chicken liver, suggesting that the enzyme is ubiquitous in mammals and birds (Barrett, 1977). Furthermore, cathepsin B-like activities have also been identified from a number of parasitic nematode and trematode species, as well as the free living nematode, *Caenorhabditis elegans* (discussed later), suggesting that such enzymes are also present in invertebrate metazoan species.

An immunohistochemical study of human tissues has demonstrated that cathepsin B is present in a wide variety of tissues and cell-types (Howie *et al.*, 1985). This study demonstrated that cathepsin B is present in cells of the lymphoid organs, the cardiovascular system, the respiratory system, the gastro-intestinal tract, the endocrine organs, the genitourinary system and the nervous system. Not all cells of these tissues and organs stained positive. The cell-types that frequently generated the strongest staining include macrophages, epithelial cells of various types, autonomous ganglion cells, parietal cells, kupffer cells, hepatocytes and neuronal cells. This study therefore clearly demonstrates that cathepsin B distribution is widespread if not ubiquitous throughout the human body.

1.7. The biological roles of cathepsin B

1.7.1. A lysosomal role for cathepsin B in protein turnover and degradation

Traditionally, cathepsin B has been considered to be a lysosomal enzyme because it was found to be greatly enriched in lysosomal fractions during purification (Barrett and Kirschke, 1981) and because cathepsin B shows optimal activity at acidic pH, consistent with a role in the acidic environment of the lysosomal compartment. This conclusion was supported by electron microscopy studies using a synthetic substrate specific to cathepsin B and antisera against the substrate which demonstrated the localisation of active cathepsin B to the lysosomes of rat liver and spleen (Barrett, 1977).

Much of the early evidence of a role for cathepsin B in turnover and degradation was obtained from studying the ability of purified lysosomal enzymes to degrade purified substrates. These results suggested that cathepsin B was capable of digesting a wide variety of different protein substrates *in vitro* including; haemoglobin, azo-casein, azo-haemoglobin (Barrett, 1977), cartilage proteoglycans (Morrison *et al.,* 1973) and collagen (Burleigh *et al.,* 1974). However, in many of these studies, the physiological relevance of the conditions was not accounted for, making it impossible to determine whether these proteins were genuine substrates of cathepsin B.

In order to overcome this problem, several laboratories performed experiments *in vivo* using protease inhibitors. Such studies were mostly performed by labelling cellular proteins by injection of ^{14}C-labelled leucine into rats, isolating rat livers, purifying the lysosomes and assaying the radioactivity of the products of protein degradation. Using such a system, Ahlberg *et al* (1985) demonstrated that lysosomes have a role in degradation of both short and long lived proteins and that leupeptin, an inhibitor of cathepsin B and other cysteine proteases, results in much reduced degradation by the lysosomes. Hopgood *et al* (1977) used a similar technique but isolated and cultured rat hepatocytes after ^{14}C-leucine labelling. The authors then added protease inhibitors, including leupeptin, to the culture media to study their effect on protein degradation. The results indicated that cysteine protease activity was important for protein degradation in these cells. Shaw and Dean (Shaw and Dean, 1980) used a similar approach but with cultured mouse peritoneal macrophages, labelling protein by addition of ^{14}C-labelled leucine into the culture media. Pepstatin was used to inhibit cathepsin D, while Z-Phe-Ala-CHN$_2$ was used to inhibit thiol proteases. Both inhibitors resulted in much reduced basal proteolysis.

Dean (Dean, 1975) studied the ability of purified cathepsin D and purified cathepsin B, as well as mixtures of lysosomal enzymes, to degrade cytosolic proteins. Again, proteins were labelled with ^{14}C-leucine by intraperitoneal injection of rats. Cytosolic proteins were extracted from homogenised rat livers and used as a substrate. Lysosomal enzyme mixtures were obtained from purified lysosomes or lysosomal-mitochondrial fractions. Cathepsin B and cathepsin D were purified from human and rat livers respectively. The results demonstrated that the rate of cytosolic protein degradation produced by the purified enzymes or lysosomal enzyme mixtures *in vitro* correlated well with rate of cytosolic protein degradation *in vivo*. This correlation suggested that lysosomal enzyme activity was important for cytosol protein degradation and that cathepsins B and D made significant contributions to this degradation.

As can be seen from the studies performed, much of the evidence for cathepsin B activity being involved in protein turnover and degradation is indirect because the inhibitors used in the studies are not completely specific for cathepsin B. Consequently, though there is good evidence to suggest that cathepsin B and other cysteine protease enzymes are important for intracellular protein turnover and degradation in the lysosomes, the relative input of cathepsin B to this process has not been determined.

1.7.2. A role for cathepsin B in the processing of protein precursors

Cathepsin B may process a number of protein precursors by limited proteolysis. Two pieces of circumstantial evidence support this. First, pairs of basic amino acid residues (which are highly susceptible to cathepsin B in synthetic substrates) have been observed to occur commonly at the point of cleavage of activation peptides from the precursors of polypeptide hormones and proteases (Steiner *et al.,* 1974). For example, the first steps of proinsulin processing seem to be cleavage at two sites which have this type of primary structure (Chance, 1972). Second, cathepsin B activity has been identified in organelles where prohormone processing is thought to occur. For example, it has been found in the purified secretion granules of a rat insulinoma and in normal rat islet granule fractions (Docherty *et al.,* 1983).

More recently, evidence of a role for cathepsin B in precursor processing comes from its own biosynthesis pathway, where procathepsin B has been shown to be capable of autoactivation by cleavage of the proregion (Mach *et al.,* 1993; Mach *et al.,* 1994a). There is also evidence to suggest that cathepsin B in conjunction with other lysosomal proteases, is responsible for the processing of thyroglobulin into active thyroxin molecules. First, thyroglobulin processing in follicular cells is thought to occur in phagolysosomes generated by the fusion of lysosomes with vesicles which transport thyroglobulin taken up from the intra-follicular space. Immunogold electron microscopy has been used to demonstrate that cathepsin B, H and L are all co-localised with Thyroxin (T4) in the colloid droplets proximal to the apical membrane within these cells (Kominami and Uchiyama, 1993). Second, selective inhibition of lysosomal proteases from purified lysosomes results in a significant reduction in thyroglobulin degradation (Dunn *et al.,* 1991b). Third, *in vitro* proteolysis of *in vivo* ^{125}I-labelled rabbit thyroglobulin by cathepsin B, L and D has shown that these enzymes are able to cleave thyroglobulin at specific sites, suggesting a direct role in processing of this hormone precursor (Dunn *et al.,* 1991a).

Cathepsin B is also thought to be the major renin processing enzyme. Crude homogenates of human kidney have been shown to contain a thiol protease which cleaves prorenin at the same site clipped in native renin (Shinagawa *et al.,* 1990). Removal of cathepsin B from such homogenates, using anti-cathepsin B antibodies or the cathepsin B specific inhibitor CA-074 abolishes this renin-processing activity (Kominami and Uchiyama, 1993). Furthermore, the authors demonstrated the co-localisation of prorenin and renin with cathepsin B in immature secretory granules using electron microscopy.

1.7.3. A role for cathepsin B in antigen processing

There is a substantial amount of evidence to suggest that cathepsin B performs important roles in the immune system. It is thought to be involved, with other lysosomal enzymes, in the digestion of endocytosed antigens for display by MHC class II molecules on the surface of antigen presenting cells (APCs). Electron microscopy studies have demonstrated the co-localisation of active cathepsin B and cathepsin D with MHC class II-I chain complexes in the early endocytic compartment, where contact with endocytosed antigens is most likely (Guagliardi *et al.,* 1990). Furthermore, processing of certain antigens have been shown to require the activity of cathepsin B *in vivo.* The immune responses of mice to vaccines of hepatitis B surface antigen (HBsAg) and rabies are inhibited by specific inhibitors of cathepsin B, specific synthetic substrates of cathepsin B or anti-cathepsin B F(ab)′ antibody fragments (Matsunaga *et al.,* 1993). In addition, using splenocytes primed with a synthetic epitope predicted to be a product of HBsAg cleavage by cathepsin B, these authors demonstrated that rechallenge with a mixture of native HBsAg and cathepsin B inhibitors produced only a weak proliferative response while rechallenge with a mixture of the synthetic epitope and cathepsin B inhibitors produced a strong proliferative response similar to native HBsAg alone. The authors concluded that this epitope was a powerful antigenic peptide and was naturally produced by cathepsin B *in vivo.*

The I chain regulates presentation of antigen by MHC class II molecules in several ways. It associates with MHC class II α,ß chains during synthesis in the endoplasmic reticulum. Subsequently, it regulates intracellular transport of MHC class II α,ß chains through cytoplasmic signals. Finally, it inhibits binding of endogenously derived peptides to MHC class II α,ß chains in living cells. Cathepsin B has been shown to digest the I chain in a staged pattern suggesting a mechanism for regulating the binding of antigenic peptides to MHC class II. (Xu *et al.,* 1994). In support of this, evidence has been obtained that cysteine proteases cleave the I chain *in vivo.* Morton *et al* (1995) studied the trafficking, proteolysis and dissociation of I chain associated with nascent MHC class II molecules in B-lymphoblastoid cells using metabolic labelling. These results demonstrated that the MHC class II-I chain complexes were catabolised rapidly after entry into lysosome-like compartments but that this catabolism was blocked by cysteine protease inhibitors.

Interestingly, gamma interferon (γ-IFN), which is known to stimulate the presentation of peptides and expression of MHC class II molecules by APCs such as macrophages and lymphocytes, has been shown to cause a selective induction of cathepsins B and L in macrophages (Lah *et al.,* 1995). The authors suggest that γ-IFN may be responsible for upregulating all genes necessary for the display of peptides by MHC class II molecules, and that cathepsins B and L are part of this pathway.

1.7.4. A role for cathepsin B in bone remodelling

Bone resorption occurs beneath the ruffled border of the mammalian osteoclast in a sealed extracellular microenvironment that resembles a phagolysosome in its acidic pH. The calcified bone is resorbed by two processes. First, acidification, which requires carbonic anhydrase activity and proton pumps, is responsible for solubilising the mineral portion of bone. Second, the organic matrix of the bone (mostly collagen) is degraded by a range of enzymes including lysosomal cysteine proteases, lysosomal hydrolases and collagenases. A role for cysteine proteases in bone resorption was first suggested after lysosomal enzymes were found to be able to degrade protein components of the extracellular matrix (see Section 1.7.1). Direct evidence for the involvement of cysteine proteases in bone resorption was subsequently obtained from experiments using specific enzyme inhibitors. Cysteine protease inhibitors were demonstrated to inhibit bone resorption by *in vitro* cultured mouse calvaria (Delaisse *et al.,* 1980). Subsequently, cysteine protease inhibitors were shown to inhibit bone resorption both *in vitro* and *in vivo* (Delaisse *et al.,* 1984), using cultured mouse calvaria and male wistar rats.

Once cysteine proteases had been implicated in bone resorption, research was directed towards determining which of the cysteine proteases might be important and how they might exert their effect. Both cathepsins B and L activities have been detected in protein extracts of bone tissue and cathepsin B activity has been detected in the media of cultured mouse calvaria (Delaisse *et al.,* 1991) suggesting these enzymes may be important effectors of bone resorption. However, cathepsin L has been demonstrated to degrade collagen and gelatin substrates at a much higher rate than cathepsin B (Delaisse *et al.,* 1991), suggesting that cathepsin B may not be a direct effector of bone resorption but rather may have a role in activating other enzymes important for this process. Recently, Hill *et al* (1994) obtained evidence for a primarily intracellular role for cathepsin B. The authors used membrane permeable and impermeable forms of two inhibitors of cysteine proteases: Ep475 which inhibits cathepsins B, H, L and S and is unable to traverse membranes;. Ep453, the ethyl ester of Ep475, which has the same specificity as Ep475 and is membrane permeable; CA-074, which specifically inhibits cathepsin B and is unable to traverse membranes; CA-074Me, the methyl ester of CA-074, which has the same specificity as CA-074 and is membrane permeable. The authors demonstrated that while Ep475, Ep453 and CA-074Me all inhibited bone resorption *in vivo* and in *vitro*, the membrane impermeable cathepsin B-specific inhibitor CA-074 had little effect on bone resorption. These results strongly suggest that while cathepsins L and S act extracellularly, and possibly intracellularly, cathepsin B acts mostly intracellularly. Accordingly, the authors suggested that cathepsin B may have a primary role in activating enzymes involved in collagen degradation such as other cysteine proteases or matrix metalloproteases..

1.7.5. A pathological role for cathepsin B in tumour cell invasion

Since biochemical studies (see Section 1.7.1) first demonstrated the ability of cathepsin B to degrade components of the extracellular matrix, this enzyme has attracted a substantial amount of work in the cancer research field because of its potential role as an effector of tumour cell invasion. These studies have resulted in the accumulation of a substantial amount of indirect evidence that suggests cathepsin B may be involved in tumour cell invasion. Elevated levels of cathepsin B activity have been correlated with increased metastatic potential or invasiveness in a variety of tumours from humans including; colorectal carcinomas (Sheahan *et al.*, 1989), neoplastic cervical cells (Pietras and Roberts, 1981) and adenocarcinomas (Recklies *et al.*, 1980). Such correlations have also been observed with tumours from rodents. For example; elevated cathepsin B activity has been observed in murine melanomas (Sloane *et al.*, 1982), while elevated cathepsin B mRNA transcript levels have been observed in murine melanomas (Qian *et al.*, 1989) and hepatomas (Moin *et al.*, 1989).

Cathepsin B from tumour cells may not only be expressed at elevated levels but also expressed as forms having larger size and/or altered stability when compared to normal tissues. For example, a study of osteoclastoma tissue revealed six cysteine protease activities (Page *et al.*, 1992). The six activities were all more similar to cathepsin B than other lysosomal cysteine proteases. All showed kinetics of inhibition by synthetic diazomethane inhibitors as expected for cathepsin B and all showed reduced activity to native collagen. Furthermore, antibodies against cathepsin B reacted against all six activities. Northern blotting using a cathepsin B cDNA probe also revealed three species of transcript for this enzyme. These results suggested that multiple forms of cathepsin B were being expressed by osteoclastomas, some of which were distinct from those observed in normal tissues.

There is a substantial amount of evidence to suggest that cathepsin B-like enzymes are secreted. Cathepsin B-like enzymes which react with anti-cathepsin B antibody and show the characteristic synthetic substrate preferences of cathepsin B have been detected in the ascitic fluid of patients with neoplasia (Mort *et al.*, 1983) and in the pleural effusions of breast cancer patients (Petrova-Skalkova *et al.*, 1987). Cathepsin B has also been detected in media conditioned by tumour cells, or explants, from a variety of different sources. Using anti-cathepsin B antibodies, cathepsin B-like enzymes have been detected in the media of cultured malignant human colorectal carcinoma cell lines (Maciewicz *et al.*, 1989) and in the media of cultured human breast tumours (Recklies *et al.*, 1982b). Using biochemical analysis, enzymes with synthetic substrate preferences and enzyme kinetics characteristic of cathepsin B have been detected in the media of cultured malignant human breast tumours (Mort *et al.*, 1980) and in the media of cultured rabbit V2 carcinoma cells (Baici and Knopfel, 1986).

In addition to being secreted, cathepsin B has also been detected in unusual cellular compartments of tumour cells. The use of synthetic substrates and synthetic inhibitors in cell fractionation experiments has demonstrated a cathepsin B-like enzyme activity on the cell surface associated with the plasma membrane of several murine melanoma and fibrosarcoma cell lines (Keren and LeGrue, 1988). Similar experiments have also identified cathepsin B-like activity strongly bound to the plasma membrane of human neoplastic cervical cancer cells (Pietras and Roberts, 1981).

Despite the large amount of indirect evidence available, there has been much debate as to the significance of the presence of membrane-associated or secreted forms of cathepsin B in tumour cells. This debate has largely arisen from two observations. First, a number of non-tumour cell lines have been shown to secrete a proportion of their cathepsin B, indicating that small amounts of cathepsin B may be present in the extracellular environment in non-disease states. For example, activatable precursor forms of cathepsin B are secreted not only by human lung tumours of different histological cell types but also by normal fibroblasts and alveolar macrophages, indicating that this process cannot be considered tumour cell specific (Werle *et al.*, 1994).

Second, most tumour cells secrete the majority of cathepsin B in an inactive precursor form. Indeed, cathepsin B activity in media from cultured tumour cells rarely shows much cathepsin B activity unless it is activated by an enzyme such as pepsin (Mort *et al.*, 1983; Baici and Knopfel, 1986; Petrova-Skalkova *et al.*, 1987; Keppler *et al.*, 1988; Werle *et al.*, 1994), suggesting that secretion of cathepsin B may not be pathologically significant.

In answer to this, many researchers have suggested that the enhanced stability at neutral and/or alkaline pH observed for a number of secreted cathepsin B-like enzymes may allow such enzymes to be active in conditions mimicking the extracellular milieu (Mort *et al.,* 1980; Recklies *et al.,* 1982a; Baici and Knopfel, 1986; Keppler *et al.,* 1988), in contrast to lysosomal cathepsin B which is known to be rapidly inactivated under alkaline conditions. More recently, Mach *et al* (1994b) have expressed recombinant human cathepsin B in yeast and isolated an active, high molecular weight form of recombinant human cathepsin B that is stable at neutral or slightly alkaline pH. This high molecular mass form was found to be a non-covalent complex between the mature recombinant cathepsin B and its propeptide. This active form was capable of hydrolysing synthetic substrates, but it was unable to hydrolyse protein substrates. However, incubation at acid pH resulted in liberation of mature active cathepsin B enzyme. This is in agreement with other studies which have demonstrated that the synthetic rat cathepsin B proregion acts as a powerful reversible inhibitor of the enzyme (Fox *et al.,* 1992), but that the inhibitory effect of this proregion is pH-dependent since acidification of the conditions causes degradation of the proregion by the enzyme itself. The authors suggested that this high molecular mass, non-covalent complex represented an important mechanism by which extracellular cathepsin B could remain dormant until local acidification allowed release of the mature active enzyme. Thus, it is possible that the cathepsin B-like enzymes secreted by some tumour cells may be activatable in the extracellular environment and may therefore be important effectors in tumour cell invasion.

1.8. Cathepsin B and parasitism

There is an increasing amount of evidence to suggest that parasitic nematode and trematode species may require cysteine proteases, including cathepsin B-like enzymes, for parasite-specific processes including parasite metabolism and host invasion.

There is a substantial amount of evidence to suggest that a cathepsin B-like enzyme is essential for *Schistosoma mansoni* metabolism. Schistosomes feed on red blood cells. These cells are lysed within the digestive tract of the parasite and haemoglobin is released. The haemoglobin is degraded in the gut producing a dark pigment end-product termed haematin. Haemoglobin digestion is therefore thought to be the principle mechanism by which schistosomes obtain essential amino acids. Two highly immunogenic proteins, named Sm31 and Sm32, have been identified in extracts from *S.mansoni* and immunofluorescence studies have localised these proteins to the cells of the digestive tract (Ruppel *et al.,* 1987). Two clones encoding Sm31 and Sm32 have been isolated from an *S.mansoni* cDNA expression library, using antisera from infected mice and humans (Klinkert *et al.,* 1987). Sequence analysis of these clones has revealed that Sm31 shares significant homology to cathepsin B. Furthermore, the expression of Sm31 in insect cells has generated a proteolytically active product capable of cleaving haemoglobin (Gotz and Klinkert, 1993). These observations suggest that the cathepsin B-like Sm31 enzyme is the 'haemoglobinase' of *S.mansoni*.

Cathepsin B-like enzymes may also perform important metabolic functions in two other trematode species, *Fasciola Hepatica* (McGinty *et al.,* 1993) and *Haplometra cylindracea* (Hawthorne *et al.,* 1993). In both cases, cathepsin B-like activities were detected in the culture media (using active site-directed affinity labels and synthetic substrates), suggesting that these enzymes may be required for digestion by these trematode species. For *F.hepatica*, a role in digestion for such enzymes is supported by the detection of cathepsin B-like enzyme activity in the secretory granules of *F.hepatica* intestinal cells (Yamasaki *et al.,* 1992). For *H.cylindracea*, which feeds exclusively on blood, the cathepsin B-like activity detected in the culture media may also be important for digestion since it was able to efficiently degrade haemoglobin (Hawthorne *et al.,* 1993).

Many parasites penetrate and migrate through host tissues during certain stages of their life cycles. It is generally believed that this must involve proteolytic degradation of the extracellular matrix. Early histochemical studies on tissue invasion by larval stages of the parasitic nematode species *Strongyloides ratti*, *Strongyloides simiae*, *Nippostrongylus muris* and *Trichinella spiralis* and the parasitic trematode species *Schistosoma mansoni* and *Schistosomatium douthitti* (Lewert and Lee, 1954) indicated disruption of host tissues around the paths of migration. The authors noted that the substantial alteration of the basement membrane and ground substances of the hosts' connective tissue was consistent with proteolytic degradation by enzymes secreted by the infective larvae. This was supported by the detection of collagenase activity from the larvae of many of the parasitic nematode and trematode species tested.

Since the early observations of Lewert and Lee, a substantial amount of indirect evidence has accumulated to suggest that many parasitic nematode and trematode species secrete proteolytic enzymes to facilitate migration through host tissues, and that some of these may be cathepsin B-like enzymes. This evidence has been obtained from the analysis of media conditioned by culture of parasitic nematode and trematode species *in vitro*. The principle problem of this approach is that it is not possible to determine whether proteolytic enzymes detected in the media have been released by active processes, such as secretion or excretion, or are a result of release of proteins after death of the parasites, which may occur as a consequence of *in vitro* culture. However, this problem has largely been resolved for a number of parasitic nematode and trematode species by reducing the *in vitro* culture time, by improving culture conditions to increase viability and by checking for worm viability prior to removing samples of the culture media. The advances made with *in vitro* culture has resulted in acceptance of the term 'Excretory / Secretory' (E/S) products to describe proteins detected in media conditioned by parasitic nematode and trematode species. However, some authors refer to these as 'In Vitro Released' (IVR) products in recognition of their uncertain origins. It is important to note that, though E/S products detected in the media are probably derived from active processes, the nature of the active processes that give rise to these products cannot be determined in such studies. Thus, E/S products detected in the culture media may be released as a result of secretion from glands, as a result of excretion from the alimentary canal, or as a result of regurgitation of enzymes from the mouth parts. Accordingly, these types of study alone cannot distinguish whether the E/S products detected in culture media may have primary roles outside the worm, suggestive of a role in host invasion or immune evasion, or primary roles within the worm, suggestive of a role in digestion.

At least one cathepsin B-like enzyme appears to be secreted from the nematode hookworm *Ancyclostoma caninum* (Harrop *et al.*, 1995). The authors isolated two cDNA clones which, though distinct from one another, both showed significant homology to human cathepsin B. The authors expressed one of these cDNA clones in *E.coli* and raised rabbit antisera to the resulting AcCP-1 recombinant protein. The antisera reacted with a single 40kDa antigen from E/S products of *A.caninum*. Immunohistochemical studies of serial sections of *A.caninum* revealed that the antisera reacted with esophageal, amphidial and excretory glands. These data provide strong evidence that a cathepsin B-like enzyme is secreted by this nematode species.

Analysis of the E/S products of the parasitic trematode *F.hepatica* using active-site-directed affinity labels has identified a major cathepsin B-like cysteine protease from both juvenile and adult worms (McGinty *et al.,* 1993). The identification of this activity in the juvenile stages, which exhibit tissue invasive behavior, suggests that this enzyme may be involved in host invasion or tissue migration. A role in tissue migration is supported by the observation that the cathepsin B-like activity detected showed unusual stability at alkaline pH and might therefore function in the extracellular environment.

The use of specific inhibitors has revealed the presence of cysteine protease activities in both the exsheathing fluid E/F and E/S products from the L_3 infective stage of the nematode *Necator americanus* (Kumar and Pritchard, 1992). Interestingly, a single cysteine protease activity was defined in the E/F fluid, but not from somatic extracts of L_3 larvae, suggesting that a single cysteine protease enzyme appears only during exsheathment of the larvae and may therefore perform a role in ecdysis. In contrast, the E/S products contained several activities that were attributable to cysteine protease enzymes, and therefore might have a role in tissue degradation during host invasion. Since only general cysteine protease inhibitors were used, it is not possible to determine whether cathepsin B-like enzymes contributed to the cysteine protease activity observed.

A cysteine protease enzyme of 28 kDa has also been partially purified from both E/S fluids and whole worm homogenates of the trematode *Spirometra mansoni* (Yong Song and Chappell, 1993). The partially purified cysteine proteases isolated from worm homogenates and from E/S fluids have identical pH profiles, both cleave collagen to produce the same products and both show negligible ability to cleave haemoglobin. Together, these data suggest that this enzyme is active both within the worm and outwith the worm. The authors concluded that the enzyme may be involved in host migration because of its ability to cleave collagen. However, a role for this enzyme in digestion is also possible since active enzyme was also present within the worm.

The E/S products generated by culture of L_2 and $L_{3/4}$ stages of the parasitic nematode *Ascaris suum* have also been studied (Knox and Kennedy, 1988). These experiments detected proteolytic activities of all four mechanistic classes. Furthermore, inhibitors specific to each of the four classes had different effects on the overall proteolysis of the selected protein substrates dependent on the pH and on which of the two life cycle stages were studied. These data suggest that both larval stages require the activity of more than one proteolytic enzyme and that these enzymes are required in a stage specific manner. A requirement for secreted enzymes from multiple mechanistic classes is probably common to all parasitic nematode and trematode species since a single enzyme class is unlikely to be able to fulfil all the biological functions performed by secreted enzymes.

Cysteine protease activities have also been identified in homogenates of a number of nematode species. For example, analysis of the E/S fluids and whole worm homogenates of both L_3 and L_4 larvae of *Dirofilaria immitis* using inhibitors to metallo and cysteine proteases has revealed that cysteine protease activity is only present in the homogenates (Richer *et al.,* 1992). Analysis of homogenates of *Strongyloides ransomi* has identified two cysteine proteases with molecular weights of 32kDa and 28kDa which show pH optima and substrate preferences similar to vertebrate cysteine proteases (Dresden *et al.,* 1985).

These studies highlight how limited our knowledge is of parasitic nematode and trematode cysteine protease biology when compared to vertebrates. The relative lack of knowledge reflects a problem faced by all parasitologists, namely the difficulties associated with growth and maintenance of parasites in the laboratory. Such difficulties limit the amount of material available for analysis and therefore hamper detailed biochemical analysis. Accordingly much of the biochemical analysis of parasite cysteine proteases in protein extracts is limited to detection and basic characterisation of activity. Despite these problems, the study of cysteine proteases in parasitic nematode and trematode species has demonstrated that some may secrete and/or excrete enzymes with cathepsin B-like activities. In turn, this suggests that the general cysteine protease activities identified in the E/S products of other parasitic nematode and trematode species may also be due in part to the presence of cathepsin B-like enzymes. These studies have also provided sufficient information to suggest that cysteine proteases, including cathepsin B-like enzymes, are used by parasitic nematode and trematode species for very different biological roles than the lysosomal cysteine proteases of vertebrates.

1.9. Cathepsin B-like multigene families

The problems associated with the study of parasites (discussed above) have resulted in the use of molecular biology as a tool for studying parasite cysteine protease biology. Accordingly, cysteine protease genes have been cloned from several unicellular and multicellular parasite species. However, to date, cathepsin B-like multigene families with the potential to encode cathepsin B-like enzymes have only been reported in parasitic nematode and trematode species. The parasitic nematode *Haemonchus contortus* possesses at least five cathepsin B-like genes (Cox *et al.,* 1990; Pratt *et al.*, 1990; Pratt *et al.,* 1992a) while the cathepsin B-like multigene family of the parasitic nematode *Ostertagia ostertagi* is thought to comprise at least two members (Pratt *et al.*, 1992b). In addition, the isolation of two cDNA clones from *Ancyclostoma caninum* that encode distinct enzymes with homology to human cathepsin B suggests this parasitic nematode species also possesses a cathepsin B-like multigene family (Harrop *et al.*, 1995). There is also some evidence for a cathepsin B-like multigene family in the parasitic trematode *Fasciola hepatica* (Heussler and Dobbelaere, 1994). The authors used polymerase chain reaction to amplify seven different cDNA clones. Five of these clones showed homology to human cathepsin L and two showed homology to *Schistosoma mansoni* Sm31 encoding a cathepsin B-like gene. However, the two cDNA clones with homology to cathepsin B were truncated in the region of the active cysteine residue and therefore it is not possible to confirm that these clones encode cathepsin B-like enzymes. The identification of cathepsin B-like multigene families in parasitic nematode species, and possibly also parasitic trematode species, suggest that this class of enzyme is important for the biology of these species. This is in agreement with the biochemical studies discussed in Section 1.8 which suggest that cathepsin B-like enzymes may be required for parasite-specific processes in some parasitic nematode and trematode species.

1.10. Cathepsin B-like enzymes in *C.elegans*

Two similar but distinct cysteine protease activities have been detected in whole worm extracts of the free living nematode species *Caenorhabditis elegans* and have been named Ce1 and Ce2 (Sarkis *et al.,* 1988b). These two activities cannot be distinguished on the basis of specificity for model substrates, cofactor requirements or inhibitor sensitivities. Both cathepsins Ce1 and Ce2 are thiol-dependent, and are both inhibited by the cysteine protease inhibitors leupeptin and E-64. They both have the same pH optima of 5.0 for the same fluorogenic synthetic substrate Z-Phe-Arg-MCA, indeed both enzymes showed maximal activity with this substrate, which is known to be susceptible to both vertebrate cathepsins B and L. However, both these enzymes possess features intermediate between cathepsins B and L. Neither of these enzymes hydrolyse Z-Arg-Arg-MCA, a characteristic substrate for cathepsin B, suggesting that these two enzymes may be more cathepsin L-like. Conversely, both enzymes display peptidyldipeptidase activity, a feature thought to be unique to cathepsin B among the papain-like cysteine protease enzymes. Thus Ce1 and Ce2 both appear to be cathepsin B-like and cathepsin L-like in their activities. These two enzymes have been shown to normally reside in lysosomes by Sarkis *et al* (1988a) which is unsurprising in view of their acidic pH optima and similarity to the lysosmal cathepsins B and L. A third cysteine protease activity, named Ce3, was also detected which was very different to Ce1 and Ce2, and indeed to vertebrate cathepsins B and L. The enzyme exhibited optimal activity at pH5.5 and was thiol dependent, yet was also leupeptin insensitive (unlike Ce1 and Ce2). This enzyme showed only very weak activity towards Z-Phe-Arg-MCA and would not hydrolyse any of the other synthetic or peptide substrates used in the study.

A gene encoding a cathepsin B-like enzyme has been cloned from *C.elegans* and named *cpr-1* (Ray and McKerrow, 1992). The predicted protein encoded by this gene appears to share the same preproenzyme structure as human cathepsin B, and shares almost 70% similarity in the mature enzyme region. The similarity of *cpr-1* to vertebrate cathepsin B suggests that these two enzymes will have a very similar structure.

However, *in-situ* hybridisation experiments suggest that *cpr-1* is expressed only in the intestinal cells lining the gut of *C.elegans*. Such tissue restricted expression is quite unlike the expression patterns observed for vertebrate cathepsin B which is thought to be expressed in most tissues of the body (Howie *et al.,* 1985). Furthermore, partial sequencing of randomly selected cDNA clones from a partially normalised *C.elegans* cDNA library identified several clones encoding partial protein sequences with homology to human cathepsin B (Waterston *et al.,* 1992). This suggests that *C.elegans* may, like the parasitic nematodes *H.contortus* and *O.ostertagi*, possess a cathepsin B-like multigene family.

1.11. Project Aims and Objectives

As a result of the extensive characterisation of vertebrate cathepsin B, the structure, function, distribution and biological roles of this enzyme are well understood. In contrast, there is only a limited amount of information available for the cathepsin B-like enzymes of parasitic nematode and trematode species. Despite this imbalance, there is sufficient evidence to suggest that cathepsin B-like enzymes perform different biological functions in parasitic nematode and trematode species to the functions performed by cathepsin B in vertebrates. These differences may make the cathepsin B-like enzymes of parasitic nematode and trematode species useful targets for serodiagnostic and/or chemotherapeutic agents. However, the effective design and application of such agents requires a sufficient knowledge of basic parasitic nematode and trematode biology and this knowledge is proving hard to obtain.

The evidence obtained to date suggests that *C.elegans* may encode a cathepsin B-like multigene family and that one member of this family exhibits gut-specific expression. These data suggest that the cathepsin B-like enzymes of *C.elegans*, like those of parasitic nematode and trematode species, may perform very different biological functions to the cathepsin B enzyme of vertebrates. Accordingly, I have initiated a study of the cathepsin B-like genes in this free living nematode species. *C.elegans* is well characterised and is a powerful tool for the isolation and analysis of new genes. Thus, the aim of this project was to isolate and characterise the cathepsin B-like genes of *C.elegans* and to gain an insight into the potential biological roles of the enzymes encoded by these genes. The information gained from these studies would not only generate more understanding of basic nematode biology but might also provide valuable insights into the function of cathepsin B-like enzymes in parasitic nematode species.

Figure 1.1

A A diagram of a cross section through the adult hermaphrodite, taken through the posterior end and viewed towards the posterior (Edwards and Wood, 1983).

Key: **m**, muscle; **h**, hypodermis; **i**, intestine; **n**, nerve cord; **g**, gonad.

B Schematic diagrams showing the lateral view of a mature adult hermaphrodite and a mature adult male (Sulston and Horvitz, 1977).

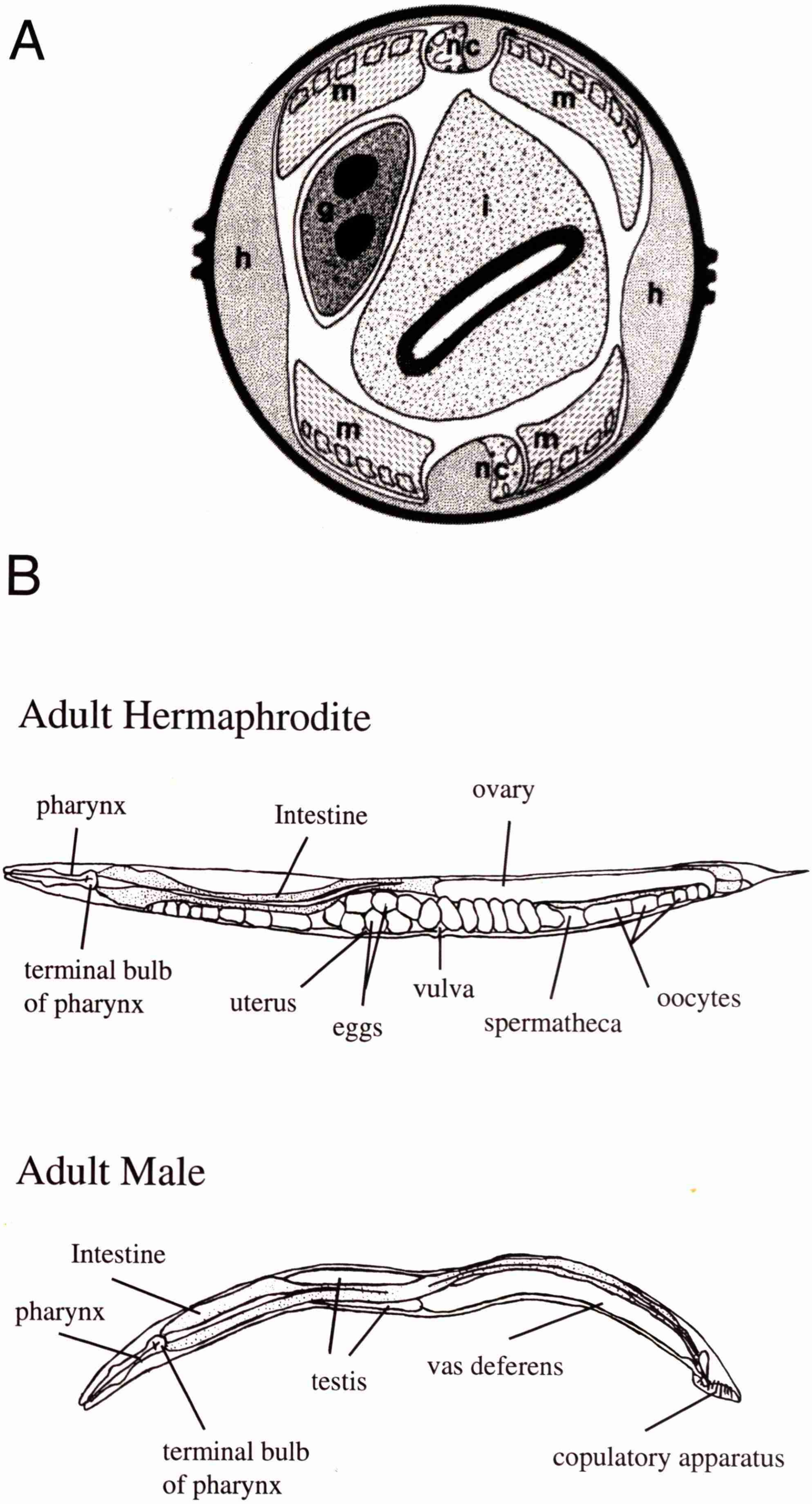
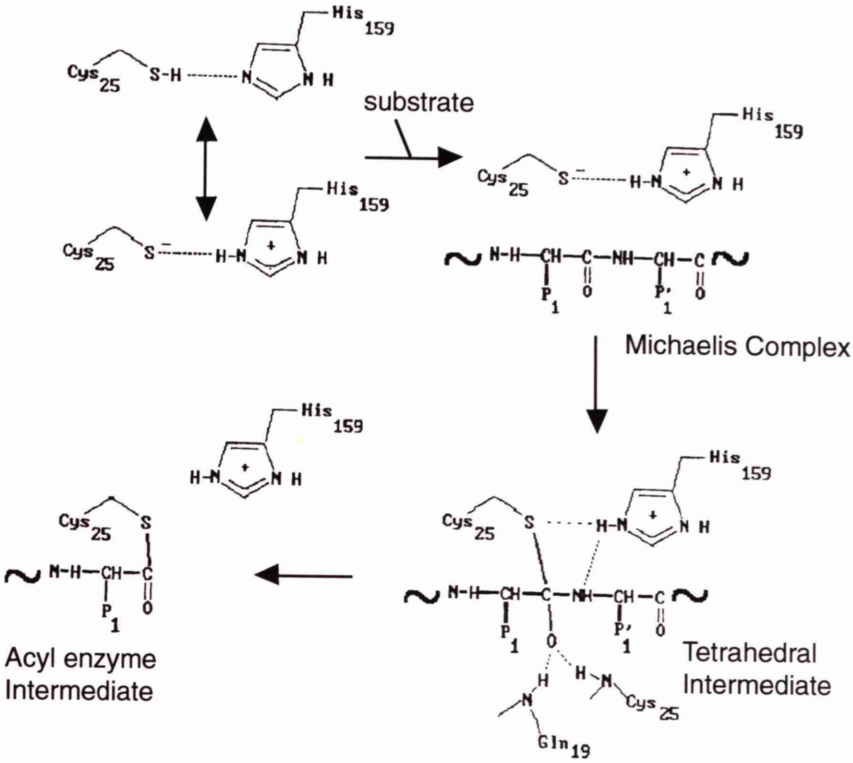

A
nc
m
m
g
i
h
h
m
m
nc
B
Adult Hermaphrodite
pharynx
Intestine
ovary
terminal bulb of pharynx
uterus
eggs
vulva
spermatheca
oocytes
Adult Male
Intestine
pharynx
testis
vas deferens
terminal bulb of pharynx
copulatory apparatus

Figure 1.2

Schematic representation of the steps involved in catalysis by members of the cysteine protease class of enzyme. This diagram shows only the reaction through to the acyl enzyme intermediate. Breakdown of this intermediate involves an enzyme-catalysed attack of water (Dunn, 1989).

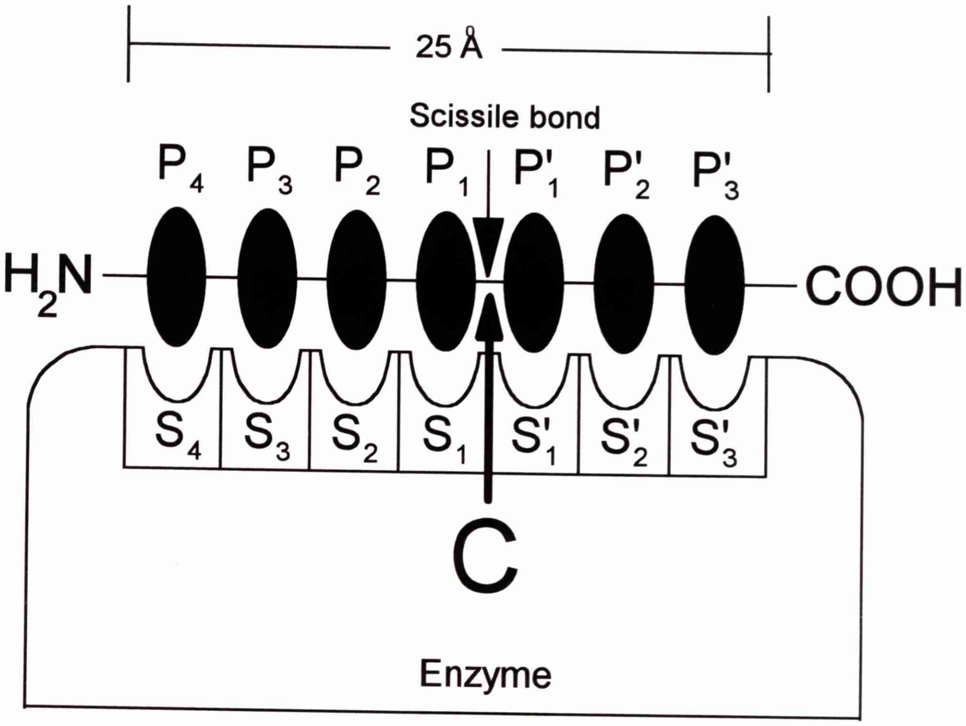


Cys25
S-H
His159
N H
substrate
S⁻
H-N
+
Michaelis Complex
N-H-CH-C-NH-CH-C
P1
O
P'1
Tetrahedral Intermediate
Gln19
Acyl enzyme Intermediate

Figure 1.3

A schematic diagram of papain showing the Schechter and Berger nomenclature for binding of a peptide substrate to papain. The seven subsites of papain are labelled S_1-S_4 and S_1'-S_3' and the amino acid residues of the peptide substrate corresponding to these sites are labelled P_1-P_4 and P_1'-P_3'. The peptide bond hydrolysed is also indicated.

The diagram was adapted from Schechter and Berger (1967).


25 Å
Scissile bond
P4
P3
P2
P1
P'1
P'2
P'3
H2N
COOH
S4
S3
S2
S1
S'1
S'2
S'3
C
Enzyme

Chapter 2

Chapter 2

Materials and Methods

Unless otherwise stated, all chemicals / reagents were purchased from BDH (Merck House, Poole, Dorset, UK), GibcoBRL (Life Technologies Ltd., Paisley, Renfrewshire, UK) or Sigma (Sigma-Aldrich Company Ltd., Poole, Dorset, UK). Restriction enzymes were obtained from GibcoBRL (Life Technologies Ltd.).

2.1. Chemical abbreviations

BSA:	bovine serum albumen
DEPC:	diethyl pyrocarbonate
diH$_2$0:	deionized water
IPTG:	isopropyl-ß-D-thiogalactoside
MOPS:	4-morpholinepropanesulphonic acid
Na$_2$EDTA:	ethylenediaminetetraacetic acid disodium salt
PEG-8000:	polyethylene glycol, molecular weight approximately 8,000
sdiH$_2$0:	sterile deionized water
SDS:	sodium dodecyl sulphate
Taurine:	2-aminoethanesulphonic acid
TEMED:	N$_1$N$_1$N$_1$N$_1$-tetramethylethylene diamine
Tris-HCl:	2-amino-2-(hydroxymethyl)-1,3-propanediol-hydrochloride
X-gal:	5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside (Boehringer Mannheim)

2.2. Commonly used stocks, solutions and media

Ampicillin (1,000x): 100mg/ml solution in sdiH$_2$O. Aliquoted and stored at -20°C.

BBL Agar: 0.6% Agar, 1% Trypticase BBL (Becton Dickinson, Oxford, UK), 0.5% NaCl, 10mM MgSO$_4$ prepared with diH$_2$O. Autoclaved and stored at room temperature.

BBL Agarose: 1% Trypticase BBL (Becton Dickinson), 0.5% NaCl, pH7.2 with NaOH, 0.65% Agarose, 0.25% $MgSO_4$ prepared with diH_2O. Autoclaved and stored at room temperature.

Denaturation Buffer: 1.5M NaCl, 0.5M NaOH in diH_2O. Stored at room temperature.

Denhardt's Solution (50x): 1% BSA. 1% Ficoll, 1% polyvinyl pyrollidone in $sdiH_2O$. Aliquoted and stored at -20°C.

DEPC Treated ddH_20: DEPC added to diH_2O to a final concentration of 0.1%, mixed vigorously for 10min and autoclaved. Stored at room temperature.

Ethidium Bromide: 10mg/ml stock in diH_2O, final concentration 0.5µg/ml in agarose gels. Stored, protected from light, at room temperature.

Herring Testes DNA: 10mg/ml stock prepared in $sdiH_2O$ and sheared using a 19 gauge hypodermic needle and lauer-lock syringe (Becton Dickinson). Stored at -20°C.

IPTG: O.5M stock in $sdiH_2O$. Aliquoted and stored at -20°C.

Kanamycin (1,000x): 50mg/ml in $sdiH_2O$. Aliquoted and stored at -20°C.

L-Broth: 1% Bacto tryptone (Difco, Michigan, USA), 0.5% yeast extract (Difco), 0.5% NaCl, in diH_2O. Autoclaved and stored at room temperature.

L-Broth Agar: as for L-Broth + 1.5g/100ml bacto-agar (Difco). Autoclaved then allowed to cool (antibiotics were added at this stage, if required) poured with sterile technique on to plastic plates, allowed to set, dried by inversion in 37°C oven and stored, for short periods, at 4°C.

M9 Buffer (1x): 0.3% KH_2PO_4, 0.6% Na_2HPO_4, 0.5% NaCl, 1mM $MgSO_4$, in diH_2O. Autoclaved and stored at room temperature.

MOPS Buffer (10x): 0.2M MOPS Buffer, 0.05M sodium acetate, 0.01M Na_2EDTA, in sdiH_2O. Stored at 4°C.

Na_2EDTA: 0.5M stock in diH_2O pH8.0 with NaOH. Autoclaved and stored at room temperature.

Neutralisation Buffer: 1.5M NaCl, 0.5M Tris-HCl pH8.0 in diH_2O. Stored at room temperature.

NGM Agar: 0.3% NaCl, 1.7% agar (Difco), 0.25% peptone (Difco), 0.0003% cholesterol (1ml/L of 5mg/ml stock in ethanol), in diH_2O. Autoclaved and then 1ml/L of 1M $CaCl_2$, 1ml/L of 1M $MgSO_4$ and 25ml/L of 1M potassium phosphate buffer pH6.0 added. Stored at room temperature.

Phenol: Purchased liquefied and pre-equilibriated. Stored at 4°C.

Phenol / Chloroform: Mix of equal volumes of phenol and chloroform. Stored at 4°C.

Potassium Phosphate Buffer (1M): Prepared from 1M KH_2PO_4 and 1M K_2HPO_4, mixed in appropriate ratio for desired pH. Autoclaved and stored at room temperature.

SDS: 10% stock solution in sdiH_2O. Stored at room temperature

Sodium Phosphate Buffer (1M): Prepared from 1M NaH_2PO_4 and 1M Na_2HPO_4, mixed in appropriate ratio for desired pH. Autoclaved and stored at room temperature.

SM Buffer: 50mM Tris-HCL (pH7.5), 0.1M NaCl, 8mM $MgSO_4$, 0.01% gelatin, in diH_2O. Autoclaved and stored at room temperature.

SSC (20x): 3M NaCl, 0.3M tri-sodium citrate. pH7.0 with NaOH, in diH$_2$O. Stored at room temperature.

SSPE (20x): 3.6M NaCl, 0.2M NaH$_2$PO$_4$, 0.02M Na$_2$EDTA pH8.0 with NaOH in diH$_2$O. Stored at room temperature.

TE: 1mM EDTA, 10mM Tris-HCl, (pH as required) in diH$_2$O. Autoclaved and stored at room temperature. TE at pH 8.0 was used for all DNA manipulations, unless otherwise indicated.

Tetracycline (1,000x): 12mg/ml in 50% sdiH$_2$O, 50% ethanol. Aliquoted and stored at 4°C.

TBE (10x): 0.9M Tris-HCl, 0.9M Boric acid, 25mM Na$_2$EDTA pH8.0, in diH$_2$O. Stored at room temperature.

TTE (20x): 1.8M Tris-HCl, 0.6M Taurine, 0.01M Na$_2$EDTA, in diH$_2$O. Stored at room temperature.

X-Gal (100x): 2% in dimethyl formamide (DMF). Stored in aliquots, protected from the light, at -20°C.

YPD Broth: 2.0% glucose, 2.0% bacto peptone (Difco) and 1% yeast extract (Difco) in diH$_2$O. Autoclaved and stored at room temperature.

YCD Broth Agar: 2% glucose, 0.17% yeast nitrogen base without amino acids (Difco), 0.5% ammonium sulphate, 1% casamino acids (Difco), 1.5% adenine sulphate, pH5.8 with acetic acid, 2% bacto agar (Difco). Autoclaved and stored at room temperature.

2xYT Broth: 1.6% Tryptone (Difco), 1% NaCl, 1% yeast extract (Difco), pH7.6 with NaOH. Autoclaved and stored at room temperature.

Superbroth: 3.5% Tryptone (Difco), 0.5% NaCl, 2% yeast extract (Difco), pH7.5 with NaOH. Autoclaved and stored at room temperature.

Note: All media and most heat insensitive solutions were autoclaved using the following conditions: 120°C, 15lbs/in^2 for 15min.

2.3. *C.elegans*, yeast and bacterial strains

C.elegans: Wild-type Bristol N2 stock (Brenner, 1974). Obtained from the *Caenorhabditis* Genetics Center (University of Minnesota, St Paul, Minnesota, USA)

Yeast: AB1380 (*MATa ψ*$^+$, *ura3*, *trp1*, *ade2-1*, *can1-100*, *lys2-1*, *his5*). Host strain for YAC clones. Gift from J.Sulston and A.Coulson.

E.coli strains:

Name	Genotype
ED8767	recA56, supE, *supF*, *hsdS*$^-$, *met*$^-$. Host cells for cosmid clones. Gift from J.Sulston and A.Coulson.
LE392	e14$^-$(McrA$^-$), *hsdR514*, *supE44*, *supF58*, *lacY1*, *galK2*, *galT22*, *metB1*, *trpR55*. Obtained from Stratagene Ltd. (Cambridge, UK).
Mc1061	*araD139*, Δ(*ara-leu*)*7696*, Δ(*lac*)*174*, *galU*, *galK*, *hsdR2*(r_{K-} m_{K+}), *mcrB1*, *rpsL*(Str^r). Gift from J.Sulston and A.Coulson.

OP50 The strain used in the our laboratory is a variant of the uracil requiring OP50 strain (Brenner, 1974) which has been transformed with a plasmid containing the Tet^{r} marker (I.L.Johnstone, pers. comm.).

Popout A popout strain of *E.coli*. The strain is a P1 lysogen and a lambda lysogen and was used for plasmid excision from lambda SHLX2 clones. Gift from J.Sulston and A.Coulson.

SURE e14^{-}(McrA^{-}), Δ(*mcr*CB-*hsd*SMR-*mrr*)171, *endA1*, *supE44*, *thi-1*, *gyrA96*, *relA1*, *lac*, *recB*, *recJ*, *sbcC*, *umuC*::Tn*5* (Kan^{r}), *uvrC*, [F′, proAB, *lacI*^{q}*Z*Δ*M15*, Tn*10* (Tet^{r})]. Obtained from Stratagene Ltd.

XL1-Blue *recA1*, *endA1*, *gyrA96*, *thi-1*, *hsdR17*, *supE44*, *relA1*, *lac* [F′, *proAB*, *lacI*^{q}*Z*Δ*M15*, Tn*10* (Tet^{r})] Obtained from Stratagene Ltd.

2.4. Vectors and clones

2.4.1. Plasmid vectors

pBluescript II KS+: pUC19 derived phagemid. Obtained from Stratagene Ltd.

pBluescript II KS-: pUC19 derived phagemid. Obtained from Stratagene Ltd.

pBluescript II SK-: pUC19 derived phagemid. Obtained from Stratagene Ltd.

pPD21.28: *lacZ* fusion vector (Fire *et al.,* 1990). Gift from A.Fire.

pPD95.03: *lacZ* fusion vector (A.Fire, pers. comm.). Gift from A.Fire.

2.4.2. Phage

VCSM13: Derived from M13 K07 mutant. Obtained From Stratagene Ltd.

lambda EMBL4 *C.elegans* Bristol N2 Genomic DNA library. Gift from I.Johnstone.

2.4.3. Clones

pCe7: The cloned insert of pCe7 is a 7kb *Bam*HI rDNA fragment containing a single repeat of the 18s and 26s ribosomal DNA sequences from *C.elegans* (Files and Hirsh, 1981). Gift from M.Krause.

pCEIF: pCEIF contains a 1615bp cDNA insert which encodes the *C.elegans* homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992). Gift from D.Roussell and K.Bennett.

pRF4: Contains the dominant *rol-6(su1006) C.elegans* cuticular collagen mutation (Mello *et al.,* 1991). Gift from J.Kramer.

cDNA Clones: cm12b6, cm14e3, cm04d10 and cm01a5 (Waterston *et al.*, 1992). Gift from J.Sulston and A.Coulson.

Cosmid Clones: ZK1037, T09D4, C29F3, K11B10, C25B8, ZK1055, T10H9, F44C4, C40C4, W02B2, W06C3, T21H3, F09F2, AD10 and F58E8. Gift from J.Sulston and A.Coulson.

YAC Clones: Y69A3, Y55B10, Y43B9, Y7E11, Y50D4 and Y48G5. Gift from J.Sulston and A.Coulson.

2.5. Culture maintenance

2.5.1. *C.elegans* culture

Bristol N2 wild-type and transgenic *C.elegans* strains were maintained at 20°C on NGM agar plates with *E.coli* OP50 as a food source according to standard methods (Sulston and Hodgkin, 1988). For large scale growth of *C.elegans*, 500ml liquid cultures were used according to standard methods (Sulston and Hodgkin, 1988). For mixed stage populations of *C.elegans*, worms were grown in 500ml liquid culture at 20°C with vigorous shaking for 3-4 days according to standard protocols (Sulston and Hodgkin, 1988). The culture was monitored at intervals by removing 1ml samples and examining the worms using a Zeiss Stemi 2000-C microscope (Carl Zeiss, Oberkochen, FRG).

2.5.2. Yeast culture

Saccharomyces cerevisiae strain AB1380 containing each of the YAC clones were maintained on YCD plates, or in 40ml of YPD media with vigorous shaking, at 30°C according to standard methods (Sherman *et al.*, 1986).

2.5.3. Bacterial culture

2.5.3.1. Bacterial culture on plates

All *E.coli* bacterial strains were grown at 37°C on L-agar plates, supplemented with the appropriate antibiotic at the appropriate concentration. The final concentrations of the antibiotics used were as follows:

i)	*E.coli* strains transformed with plasmids:	100µg/ml ampicillin or 12µg/ml tetracycline
ii)	*E.coli* strains transformed with cosmids:	75µg/ml ampicillin or 30µg/ml kanamycin

2.5.3.2. Bacterial liquid culture

Unless otherwise stated, all liquid cultures were grown at 37°C with vigorous shaking. The media, antibiotics and antibiotic concentrations used in these cultures were as follows:

i) *E.coli* strains transformed with plasmids:
L-broth supplemented with 100μg/ml ampicillin or 12μg/ml tetracycline.

ii) *E.coli* strains transformed with cosmids:
2xYT supplemented with 120μg/ml ampicillin or 75μg/ml kanamycin.

2.6. DNA preparation

2.6.1. Plasmid DNA mini preparation

Magic and Wizard Miniprep kits (Promega Corporation, Southampton, UK), or QIAprep-spin kit (Qiagen Inc., via Hybaid Ltd., Teddington, Middlesex, UK), were used for plasmid DNA mini preparations from an appropriate volume of overnight culture of *E.coli* cells, according to the manufacturer's protocols supplied with these kits. Alternatively, plasmid DNA was isolated from 1.5ml of overnight cultures of *E.coli* cells by alkaline extraction (Birnboim and Doly, 1979), organic extraction and ethanol precipitation according to standard protocols (Sambrook *et al.*, 1989). The plasmid DNA was resuspended in an appropriate volume of TE or sterile distilled water, usually 50μl.

2.6.2. ssDNA template preparation for sequencing

i) 1ml of overnight culture of *E.coli* cells containing the appropriate plasmid was used to inoculate 30ml of superbroth and grown at 37°C for 75min in a shaking incubator.

ii) 30μl of VCSM13 helper phage with a titre of approximately 1x10^{11} pfu/ml (Stratagene Ltd.) were added to the culture and incubated in the same conditions for a further 7 hours.

iii) The culture was centrifuged at 10,000 rpm for 10min using a Beckman J2.HS centrifuge and JA-17 rotor (Beckman Instruments Inc., Palo Alto, California, USA).

iv) Phage particles containing the ssDNA template were precipitated using 0.25 volumes of 20% PEG-8000, 3.5M ammonium acetate at room temperature for 15min.

v) The precipitated phage particles were centrifuged 10,000 rpm for 10min as above.

vi) The phage particles were resuspended in 1ml of TE and reprecipitated as above.

vii) The pelleted phage particles were resuspended in 700μl of TE, phenol extracted (3 times), Phenol / chloroform extracted (1 time), chloroform extracted (1 time), ethanol precipitated, washed and resuspended in 15μl of TE using standard protocols (Sambrook *et al.,* 1989).

2.6.3. Plasmid DNA medium and large scale preparations

Qiagen Plasmid Midi and Plasmid Maxi DNA Preparation kits (Qiagen Inc.) were used for medium and large scale plasmid DNA preparations respectively, from an appropriate volume of overnight culture of *E.coli* cells according to the manufacturer's protocols.

2.6.4. Large scale cosmid DNA preparations

Large scale cosmid DNA preparations were made from 200-250ml cultures of *E.coli* cells containing the desired cosmid. The cultures were grown in two phases; a 20ml overnight culture was prepared and 10ml of this used to innoculate 200-250ml of fresh media which was grown to late log phase. Cosmid DNA was extracted using the alkaline hydrolysis method (Birnboim and Doly, 1979; Sambrook *et al.,* 1989) with minor modifications communicated by Alan Coulson. The modifications are listed below:

i) Solution I was prepared with Tris-HCl pH7.4 and 8ml of this solution was used.

ii) No lysozyme step was used.

iii) 8ml of Solution II was used.

iv) Solution III was replaced with 3M Sodium Acetate pH4.8, 6ml were used and the mixture was left on ice for 30min.

v) Centrifugation of the precipitate generated by addition of 3M Sodium Acetate pH4.8 was performed at 12,000 rpm for 10min

vi) The isopropanol precipitation and phenol/chloroform steps were omitted.

vii) The precipitation step using 2 volumes of ethanol was performed at room temperature for 5min, and no salt was added.

iix) The cosmid DNA pellet was resuspended in 2.5ml of 10mM Tris-HCl pH7.4, 10mM Na_2EDTA overnight at 4°C.

Solution I :	50mM glucose, 25mM Tris-HCl (pH8.0), 10mM Na_2EDTA, 2mg/ml lysozyme added fresh prior to use.
Solution II :	0.2M NaOH, 1% SDS.
Solution III:	3M sodium acetate (pH4.8).

The cosmid DNA was purified using continuous caesium chloride / ethidium bromide density gradient centrifugation, isobutanol extraction and ethanol precipitation according to standard protocols (Sambrook *et al.*, 1989).

2.6.5. Lambda phage DNA preparations

Lambda phage DNA was isolated and purified using the Promega Magic Lambda Prep kit (Promega Corporation) according to the manufacturer's protocol and using *E.coli* LE392 as the host strain. The *E.coli* LE392 host cells were cultured according to the liquid culture method of the protocol.

2.6.6. Yeast genomic DNA preparation

40 ml of yeast containing the appropriate yeast artificial chromosome (YAC) clones were grown in YPD media for 48 hours at 30°C. High molecular weight DNA (Genomic DNA and YAC DNA) was extracted and purified according to the yeast DNA miniprep protocol of Sherman *et al* (1986) with one alteration: the zymolyase protoplasting step was replaced by treatment with 10U of lyticase at 37°C for 4 hours.

2.7. RNA preparation

2.7.1. Isolation of mixed stage embryos

Worms were grown in 500ml liquid culture at 20°C with vigorous shaking until the *E.coli* OP50 food source was just clearing (approximately 4 days) according to standard protocols (Sulston and Hodgkin, 1988). The worms were purified from contaminating bacteria by sucrose floatation according to standard methods (Sulston and Hodgkin, 1988) with one exception: the worms were not left to settle overnight at 4°C prior to the sucrose floatation step. Eggs were isolated by treatment of the worms with hypochlorite according to standard protocols (Sulston and Hodgkin, 1988).

2.7.2. RNA extraction

RNA was extracted and purified from eggs (see Section 2.7.1) and mixed stage populations of *C.elegans* (see Section 2.5.1). RNA was extracted from eggs or mixed stage worms as follows:

i) Worms or eggs were ground in liquid nitrogen using a pestle and mortar.

ii) The ground material was added to a solution of 4M guanidinium thiocyanate, 0.13% sarkosyl, 33mM Tris-HCl pH 8.0, 0.5% ß-mercaptoethanol and 6.7mM Na_2EDTA and homogenised thoroughly using a PTFE hand homogeniser.

iii) The solution was phenol (pH4.3) / chloroform extracted three times according to standard protocols (Sambrook *et al.*, 1989).

iv) The RNA was precipitated with 0.3M sodium acetate and 0.75 volumes of ethanol

v) The pelleted RNA was resuspended in 10ml of the guanidinium thiocyanate solution described above, reprecipitated with 0.5 volumes of ethanol, centrifuged, washed and resuspended in 50-200µl of DEPC treated sterile distilled water according to standard methods (Sambrook *et al.*, 1989).

Poly(A)$^+$ selected RNA was obtained using the PolyATtract kit (Promega Corporation) according to the manufacturer's protocol.

2.8. Standard molecular biology techniques

2.8.1. Restriction endonuclease digests

Single enzyme digests were carried out at 37°C in 1x enzyme reaction buffer supplied with the restriction endonuclease. Double and triple digests were performed in a buffer which yielded at least 50% activity for each of the enzymes used. There were situations where the enzymes used required incompatible reaction conditions. In these cases heat inactivatable enzymes were used first, inactivated by heating at 65°C for 10min, the DNA precipitated, washed with ethanol and resuspended in sterile distilled water containing 1x reaction buffer appropriate for the next restriction endonuclease to be used. In other cases, the restriction enzymes could not be heat inactivated and therefore this step was replaced with organic extraction. The quantity of enzymes used varied depending upon the source of the DNA and also upon its required usage, however a general rule was observed which states that 1 unit of enzyme digests 1μg of lambda DNA in 1 hour in a 20μl volume at the appropriate temperature. For genomic DNA to be used for Southern analysis or DNA to be used for preparation of subclones, a 5x excess of enzyme was used to ensure complete digestion of the DNA.

2.8.2. Agarose Gel electrophoresis of RNA and DNA

2.8.2.1. Agarose gels for electrophoresis of DNA samples

Restriction endonuclease digested DNA and PCR amplified DNA samples were electrophoresed through agarose gels. Gels ranging from 0.6% to 2% agarose were prepared by dissolving the appropriate amount of agarose in 1x TBE and adding 0.2μg/ml ethidium bromide. These gels were run in 1x TBE using gel electrophoresis tanks from GibcoBRL or Pharmacia (Milton Keynes, UK). The gels were generally run at a voltage equivalent to 3V/cm. All DNA samples were loaded into the wells of these gels using DNA sample buffer (see below). The electrophoresed DNA was visualised using a UV transilluminator (Appligene, Durham, UK) and a record of the gel made using a Polaroid camera or a digital imaging system (Appligene). Each gel contained at least one lane of DNA size standards, either 1 kb ladder or lambda/*Hin*dIII (both from GibcoBRL)

Fragment Sizes (in kilobases)

1 kb ladder	12.216	11.198	10.180	9.162	8.144
	7.126	6.108	5.090	4.072	3.054
	2.036	1.636	1.018	0.517	0.506
	0.396	0.344	0.298	0.220	0.201
	0.154	0.134	0.075		
lambda / *Hin*dIII	23.130	9.416	6.557	4.361	2.322
	2.027	0.564			

10x DNA Sample buffer

50% glycerol
0.5% bromophenol blue
0.5% xylene cyanol
in 1X TBE

2.8.2.2. Agarose gels for electrophoresis of RNA samples

Denaturing 1% agarose gels were prepared in 1x MOPS running buffer and 2% formaldehyde (final concentration) according to the Quick Formaldehyde RNA Gel protocol of the ZAP-cDNA kit (Stratagene Ltd.). These gels were run in 1x MOPS running buffer using gel electrophoresis tanks from GibcoBRL. According to the Stratagene Quick Formaldehyde RNA Gel protocol, RNA samples (final volume 2µl) were mixed with 10µl of RNA sample buffer (see below) and boiled for 2min prior to loading. For total RNA samples, ethidium bromide was included in the sample buffer. For poly(A)$^+$ selected RNA samples, the ethidium bromide was omitted. Each gel contained one lane of RNA size standards (GibcoBRL). The electrophoresed RNA was visualised using a UV transilluminator (Appligene) and a record of the gel made using either a Polaroid camera or an digital imaging system (Appligene).

Fragment Sizes (in kilobases)

0.24-9.5 kb RNA Ladder	9.49	7.46	4.40	2.37
	1.35	0.24		

RNA sample buffer

48% formamide
1x MOPS running buffer
6.4% formaldehyde
5.3% glycerol
5.3% of a saturated bromophenol blue solution
666µg/ml ethidium bromide (optional)

2.8.3. Gel purification

Gel purification of DNA fragments generated from restriction endonuclease digests or PCR reactions were performed in two ways. All fragments were electrophoresed through agarose gels run in 1x TBE. The appropriate fragments were excised using a scalpel blade and the DNA either extracted from the gel using the QIAEX DNA or QIAquick Gel Extraction kits (both from QIAGEN Inc.) according to the manufacturer's protocols.

2.8.4. Transformation of *E.coli*

Transformations were performed using *E.coli* XL1-blue supercompetent cells and 25-100ng of ligated DNA according to the manufacturers protocol with the following exception: Only 20µl of supercompetent cells were used for the transformations and the volumes of all solutions required for the transformation were adjusted accordingly. Alternatively, competent *E.coli* cells were prepared by treatment with 50mM $CaCl_2$ and transformed with plasmid DNA according to standard protocols (Sambrook *et al.*, 1989).

The transformed cells were selected using the antibiotic resistance marker carried by the plasmid (Amp^r). Where applicable, blue / white colony selection was also used to select transformants with plasmids containing inserts. This selection makes use of α-complementation of the *lacZΔM15* mutation in appropriate *E.coli* strains. In these instances, the agar media was also supplemented with 0.004% X-gal and 34µM IPTG (final concentrations).

2.8.5. Ligation reactions

Ligation reactions were performed in a 10μl or 15μl volume using 1x T4 DNA ligase buffer (Promega Corporation) and 1-2 units of T4 DNA ligase (Promega Corporation). Generally, 250ng of vector DNA were used in each ligation. The amount of insert DNA used varied with the molar concentration of free ends in the reaction. Ligations were performed overnight either at 16°C or at room temperature.

2.8.6. Southern and Northern blots

Northern blots and Southern blots were prepared from RNA and DNA electrophoresed through the appropriate type of agarose gel (see Section 2.8.2). All Northern and Southern Blots were performed using Hybond N (Amersham International plc, Little Chalfont, Buckinghamshire, UK) according to the manufacturers protocols. Southern blots were performed for 24 - 48 hours, Northern blots for 24 hours. The Southern blots were rinsed in 2x SSC and either baked at 80°C for 2 hours or UV cross-linked. The 2x SSC wash step was omitted for Northern blots and all Northern blots were UV cross-linked. All UV cross-linking was performed using a D-1000 cross-linker (Ultraviolet Products Ltd., Cambridge, UK).

2.8.7. Probe preparation

2.8.7.1. Random priming

All radiolabelled dsDNA probes were prepared by random priming using 50-100ng of DNA, ($\alpha-^{32}$P)dCTP (Dupont UK Ltd., Stevenage, UK) and the Prime-It or Prime-it II Kits (Stratagene Ltd.) according to the manufacturer's protocols. All cloned cDNA probes were gel purified from their vectors prior to labelling. All PCR amplified probes were gel purified from other amplified bands prior to labelling.

2.8.7.2. 5′ end labelling

Oligonucleotides (200ng) were end labelled using (γ-^{32}P)ATP (Dupont UK Ltd), 10U of polynucleotide kinase (Promega Corporation) and 1x polynucleotide kinase buffer supplied with the enzyme in a final reaction volume of 10μl. The kinase reaction was allowed to proceed at 37°C for 45min.

2.8.7.3. Riboprobes

Antisense riboprobes were prepared using the T3 or T7 RNA polymerases, (γ-^{32}P)rUTP (Dupont) and the RNA Transcription Kit (Stratagene Ltd.) according to the manufacturer's protocols but with the addition of RNasin Ribonuclease inhibitor (Promega Corporation).

2.8.7.4. Probe purification

All probes were purified using NucTrap Probe Purification Columns (Stratagene Ltd.) according to the manufacturer's protocols.

2.8.8. Hybridisation

All hybridisations were performed in Hybaid (Hybaid Ltd., Teddington, Middlesex, UK) hybridisation bottles using Bachofer (Reutlingen, FRG) hybridisation ovens. All Northern and Southern blots were stripped using a boiling solution of 0.1% SDS after exposure to medical X-ray film.

2.8.8.1. Southern hybridisations

All Southern hybridisations were performed according to the Amersham Hybond N protocols (Amersham International plc) with the following alterations:

i) The solution used for prehybridisation and hybridisation was altered to:

6x SSC
5x Denhardt's Solution
0.5% SDS
20μg/ml Herring testes DNA (Denatured by heating at 100°C for 10min prior to addition to the solution)

ii) Prehybridisations were performed for at least 3 hours using 25ml of the above solution

iii) Prior to addition of the radiolabelled probe, sufficient prehybridisation solution was removed such that the remainder barely covered the Southern blots.

All radiolabelled, random primed dsDNA probes were denatured by heating at 100°C for 5min prior to addition to the hybridisation bottles. All Southern hybridisations were performed for at least 16 hours at 65°C unless 5′ end-labelled oligonucleotide probes were being used, in which case the hybridisations were performed for at least 16 hours at 55°C.

2.8.8.2. Northern hybridisations

The Northern blots were rolled up and inserted, RNA side up, into the hybridisation bottles. The Northern blots were prehybridised and hybridised in the same solution. The prehybridisation / hybridisation solution comprised, 50% formamide, 5x SSPE, 5x Denhardts, 0.5% SDS and 100μg/ml herring testes DNA. Prehybridisations were performed for at least 4 hours and hybridisations were performed for at least 16 hours. Both prehybridisation and hybridisation were performed at 42°C. Prior to addition of the radiolabelled probe, sufficient prehybridisation solution was removed such that the remainder barely covered the Northern blots. Random primed cDNA probes were denatured at 100°C for 10min and riboprobes were heated at 65°C for 5min prior to addition to the hybridisation solution.

2.8.8.3. Southern hybridisation washes

All hybridised Southern blots were washed once at low stringency in at least 1 litre of 2x SSC / 0.1% SDS solution at room temperature for 30min All Southern blots probed with random primed DNA probes were washed with two high stringency washes using at least 1 litre of 0.1x SSC / 0.1% SDS solution at 65°C for 15min. For Southern blots probed with 5′ end-labelled oligonucleotide probes, the two high stringency washes were replaced with a single medium stringency wash performed using at least 1 litre of 0.5x SSC / 0.1% SDS solution at 55°C for 15min.

2.8.8.4. Northern hybridisation washes

Both total RNA Northern blots probed with random primed cDNA probes and poly(A)$^+$ RNA Northern blots probed with riboprobes were initially washed using at least 1 litre of 2x SSPE / 0.1% SDS solution at room temperature for 20min. All Northern filters were subsequently washed at medium stringency using at least 1 litre of 1x SSPE / 0.1% SDS solution at 65°C for 15min. A high stringency wash was also performed where required, using at least 1 litre of 0.1x SSPE / 0.1% SDS solution at 65°C for 10min.

2.8.8.5. Autoradiography

Probed Southern and Northern blots were sealed in bags and exposed to sheets of Agfa Curix RP1 medical X-ray film (via H.A.West, Clydebank, Dumbartonshire, UK) in autoradiography cassettes with intensifying screens. Exposures were carried out at -70°C until satisfactory images were produced. The autoradiography film was developed using an M35-M X-omat processor (Kodak Ltd., Hemel Hempstead, Hertfordshire, UK).

2.8.8.6. Colony hybridisation

Colony hybridisation was performed using replica plates. Each colony to be screened was streaked onto two L-agar plates, both supplemented with the appropriate antibiotic and one overlaid with an 88mm Hybond N (Amersham International plc) Nylon filter. A template was used to streak each colony onto the surface of the two plates in an ordered grid. Thus, each colony was first streaked onto the master plate at a position defined by the template. The colony was subsequently streaked onto the Hybond N filter (Amersham International plc) of the second plate at the equivalent position. The master and replica plates were grown overnight at 37°C. The bacterial colonies on the Nylon filter were lysed and the DNA was denatured and neutralised according to standard protocols for colony hybridisation (Sambrook *et al.,* 1989). The DNA was cross-linked to the filters by baking for 2 hours at 80°C. The resulting filters were hybridised according to the conditions described for Southern hybridisation (Section 2.8.8.1) using random-primed cDNA probes (Section 2.8.7.1) and were washed twice at high stringency (Section 2.8.8.3) prior to autoradiography (Section 2.8.8.5).

2.9. Screening a lambda EMBL4 *C.elegans* Bristol N2 genomic DNA library

E.coli LE392 was used as a host strain for phage of the lambda EMBL4 library. The cells were cultured in L-broth supplemented with 0.2% glucose and 10mM $MgSO_4$ at 37°C overnight The cells were centrifuged at 3,000 rpm using a Beckman J2-HS centrifuge and JA-17 rotor (Beckman Instruments Inc.). The pelleted cells were resuspended in 10mM $MgSO_4$ to generate a suspension with an OD_{600} of 2.

2.9.1. Determining the titre of the lambda library

The titre of the lambda EMBL4 library (prepared by Dr I.L.Johnstone) was determined using standard methods (Sambrook *et al.,* 1989) with the following exceptions:

i) 100µl of *E.coli* LE392 host cells were transfected with 10µl of each serial dilution (1:1-1:10,000) of lysate from the amplified library.

ii) The infected cells were plated onto BBL bottom agar, in 9cm petri dishes, using 3ml of BBL top agarose.

2.9.2. Plating and screening of the lambda library

Approximately 90,000 lambda phage from the lambda EMBL4 library (15,000/plate) were plated and screened according to standard methods (Sambrook *et al.*, 1989), with the following exceptions:

i) Initially, 50μl of phage library lysate (approximately 15,000 phage) were mixed with 100μl of *E.coli* LE392 host cells

ii) The infected cells were plated onto BBL bottom agar, in 9cm petri dishes, using 3ml of BBL top agarose.

iii) Hybond N nylon filters (Amersham International plc) were used for the transfer of phage DNA.

iv) After denaturation and neutralisation, the filters were baked at 80°C for 2 hours to fix the DNA to the filters.

v) The immobilised DNA was hybridised with a ^{32}P-labelled cm12b6 cDNA probe (radiolabelled by random priming, Section 2.8.7.1) using the conditions described in Section 2.8.8.1. The filters were washed twice at high stringency (Section 2.8.8.3) prior to autoradiography (Section 2.8.8.5).

2.10. Excision of cDNA clones from lambda SHLX2

The pRATII plasmids containing the cDNA clones were excised from lambda SHLX2 phage according to the instructions supplied with the phage clones. Lysates from the lambda SHLX2 cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 (Waterston *et al.*, 1992) were obtained from J.Sulston and A.Coulson. *E.coli* MC1061 cells were grown at 37°C with vigorous shaking until late log phase. The cells were pelleted and resuspended in 10mM $MgSO_4$ to generate a suspension with an OD_{600} of 1 (as described in Section 2.9). 100μl of the cell suspension were transfected with 1μl of 10-fold serial dilutions of lysate (1:1-1:1,000) from each of the four cDNA clones and incubated at 37°C for 15min. The infected cells were plated using standard methods (Sambrook *et al.*, 1989), except that 3ml of BBL top agarose were used to plate the infected cells onto BBL bottom agar, in 9cm petri dishes. The plates were incubated at 37°C overnight. Agarose plugs from single well isolated plaques for each of the cDNA clones were obtained, eluted in 1ml of SM buffer and stored at 4°C with 10μl of chloroform according to standard methods (Sambrook *et al.*, 1989). 10μl of each resulting plaque lysate were added to 100μl of an overnight culture of a popout strain of *E.coli* (P1 and lambda lysogen) and 10μl of VCSM13 helper phage (titre: 6.1×10^{11} pfu/ml), incubated at room temperature for 75min, added to 2ml of 2xYT supplemented with ampicillin (90μg/ml) and incubated at 37°C overnight. The overnight culture was centrifuged and 5μl of the resulting supernatant mixed with 100μl of an overnight culture of *E.coli* XL1-blue, incubated at room temperature for 75min, streaked onto L-agar plates supplemented with ampicillin (90μg/ml) and tetracycline (15μg/ml) and grown at 37°C overnight. Large isolated colonies were selected for further analysis. The cDNA inserts were subsequently subcloned as *Apa*I/*Sac*I fragments into pBluescript II KS+ for further analysis.

2.11. Nucleotide sequencing

All sequencing was performed using the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) using the Sequenase version 2.0 kit or the Sequenase version 2.0 Quick Denature kits and (α-^{35}S)dATP (Dupont UK Ltd.) according to the manufacturer's protocols (Amersham International plc). The sequencing reactions were performed using single or double stranded DNA templates and either custom made sequencing primers or primers that anneal to sequences flanking the pBluescript polylinker, into which the cDNA and genomic clones were inserted. The sequencing reactions were electrophoresed at 70 watts through a denaturing 6% acrylamide gel (see below) with 1xTBE using Model S2 (GibcoBRL) or Base Ace sequencing gel tanks (Stratagene Ltd.). The sequencing gels were fixed in 10% methanol, 5% glacial acetic acid at room temperature for 30min. The gels were dried onto 3MM paper at 80°C for 2 hours using a Bio-Rad Model 583 gel drier (Bio-Rad Laboratories Ltd., Hemel Hempstead, Hertfordshire, UK). The dried gels were exposed directly to sheets of Agfa

Curix RP1 medical X-ray film (via H.A.West) in autoradiography cassettes with intensifying screens, exposures were carried out at room temperature until satisfactory images were produced. The autoradiography film was developed using an M35-M X-omat processor (Kodak Ltd.).

Sequencing gel:

50% UREA
6% acrylamide (19:1 Acrylamide/Bis-acrylamide,
Anachem, Luton, Bedfordshire, UK)
1xTBE (replace with 1xTTE for glycerol tolerant gels)
0.1% ammonium persulphate
(polymerisation of acrylamide effected by addition of TEMED to a final
concentration of 0.01%).

2.11.1. Preparation of custom made primers

Custom oligonucleotide primers for nucleotide sequencing and Polymerase Chain Reaction (PCR) were obtained from two sources. They were either synthesised by myself using a PCR Mate DNA synthesiser (Applied Biosystems, Warrington, Cheshire, UK) or were supplied ready to use by Strathclyde University (Glasgow, UK).

2.11.2. Deprotection of synthesised oligonucleotides

The protective group of primers synthesised using the Applied Biosystems PCR Mate DNA synthesiser must be removed prior to use. The primers were deprotected and purified using the following protocol:

i) The beads supporting the synthesised oligonucleotides were transferred from the column into a 1.5ml screw cap eppendorf; 1ml of ammonium hydroxide (30% w/v) was added and the tube was incubated at room temperature for 2 hours with occasional vortexing.

ii) The eppendorfs were centrifuged at 15,000 rpm for 5 minute in a microfuge (Heraeus Equipment Ltd., Brentwood, Essex, UK), the supernatant was transferred to a new screw cap eppendorf, ammonium hydroxide (30% w/v) was added to the top and the tube was incubated overnight at 55°C.

iii) The solution was transferred to a 12ml polypropylene centrifuge tube and precipitated using 0.1 volumes of 5M ammonium acetate and 3 volumes of ethanol for 1 hour at -20°C.

iv) The precipitated DNA was pelleted by centrifuging at 6,000 rpm for 10min using a J2-HS centrifuge and JA-17 rotor (Beckman Instruments plc)

v) The oligonucleotides were resuspended in 500µl of TE, precipitated with 0.1 volumes of 3M sodium acetate pH5.2 and 3 volumes of ethanol for 1 hour at 20°C and centrifuged as above.

vi) The oligonucleotides were washed in 70% ethanol, dried and resuspended in 200-500µl of TE.

2.11.3. Nucleotide sequence of the primers used for sequencing

2.11.3.1. Primers for sequencing the 5′ rapid amplification of cDNA ends (RACE) products

All the cloned 5′ RACE products were sequenced using primers flanking the pBluescript II KS+ vector polylinker. The primers were all obtained from Stratagene Ltd. and are listed below:

Name	Primer sequence
T3 17mer	5′-GAAATCACTCCCAATTA-3′
T7 17mer	5′-AATACGACTCACTATAG-3′
sk 17mer	5′-CTAGGTGATCAAGATCT-3′
ks 17mer	5′-CGAGGTCGACGGTATCG-3′

2.11.3.2. Primers for sequencing the cDNA and genomic clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The nucleotide sequences of the cDNA and genomic clones were obtained using a combination of primers flanking the pBluescript vector polylinker and custom made primers. Most of the genomic clone nucleotide sequences were obtained using custom made primers, however some sequence data were obtained using the T7 17mer whose nucleotide sequence is given above. Most of the nucleotide sequence of the cDNA clones was also obtained using custom primers however some sequence data were also obtained using the T3 17mer (listed above) and the M13 (-40 Forward) primer supplied with the Sequenase kits (Amersham International plc). The nucleotide sequence of this primer is listed below:

Name	Primer sequence
M13 (-40 Forward)	5′-GTTTTCCCAGTCACGACGTTGTA-3′

The nucleotide sequences of the custom made primers used to obtain the majority of the nucleotide sequences for the cDNA and genomic clones of each of the four genes are listed below. The position of the primers with respect to the ATG initiation codons of each of the four genes is also given to help understand how the final sequence of each of the four genes (Chapter 4, Figures 4.11, 4.12, 4.13 and 4.14) was obtained.

Table i. Custom primers for *cpr-3*

Name	Primer (Read 5′-3′)	Clones Sequenced With Primer	Sense of Primer	Position of 5′ End and 3′ End Bases of Each Primer With Respect to ATG Initiation Codon	
				5′ End	3′ End
cpr-3/0.5	AGTTTTCAGTTTTCCCTC	genomic	sense	-46	-29
cpr-3/1	CCGCGGAATCAATATTG GGCAGTC	cDNA and genomic	sense	+114	+137
cpr-3/2	CACCGTACAAACCAGCT GGGTGGC	cDNA and genomic	sense	+168	+191
cpr-3/3	TACGGATGTAAGGGAGG C	cDNA and genomic	sense	+526	+543
*cpr-3*int/2.5	ACGAGAATTTTTATCAGA TG	genomic	sense	+806	+825
cpr-3/4	AATCTACCATTACGGAC CTGTG	cDNA and genomic	sense	+1198	+1219
cpr-3/5	ACTATACTTCCGGAAAG C	cDNA and genomic	sense	+1275	+1292
cpr-3/6	GAGTGCTCTGATGTATG	cDNA and genomic	sense	+1671	+1687
cpr-3/R0.5	GAAAACTGGAAAGAAGC G	genomic	antisense	-40	-57
cpr-3/R1	GC**GTCGAC**TGGTTTGTA CGGTGTTCACAT	cDNA and genomic	antisense	+182	+161
cpr-3/R2	TCACAGGTTGTTGAGTT CCATT	cDNA and genomic	antisense	+481	+459
*cpr-3*int/R2.5	CTTGAACTCGTGACCTC C	genomic	antisense	+944	+927
cpr-3/R3	CGGTGACGGATTTCGTA G	cDNA and genomic	antisense	+1184	+1167
cpr-3/R4	CAATTCCAGCTACCACG	cDNA and genomic	antisense	+1460	+1444
cpr-3/R5	CGTCAGCACCATCAATG T	genomic	antisense	+1596	+1579
cpr-/5′FL	GTACAATTGCATAAATCT C	genomic	sense	see below	see below
cpr-/3′FL	GAACTTCACTGATAATAC	genomic	antisense	see below	see below

Table ii. Custom primers for *cpr-4*

Name	Primer (Read 5′-3′)	Clones Sequenced With Primer	Sense of Primer	Position of 5′ End and 3′ End Bases of Each Primer With Respect to ATG Initiation Codon	
				5′ End	3′ End
cpr-4/0.5	TTGCTCTATCTTGCTATTTGCTCTT	genomic	sense	-33	-9
cpr-4/1	GCCTTGGTGGCTGTTACCGCCGG	cDNA and genomic	sense	+71	+93
cpr-4/2	CTATCACCGAGTATGTGAACTCAAAGCAA	cDNA and genomic	sense	+129	+157
cpr-4/3	TCTTGCTGCTCCAACTGCG	cDNA and genomic	sense	+470	+488
cpr-4/4	CACCGCTTACGCCGTCGGAAAG	cDNA and genomic	sense	+739	+760
cpr-4/5	GGAACCAACGAGTGCGG	genomic	sense	+1003	+1017
cpr-4/R0.5	AAAAGAGCAAATAGCAAG	genomic	antisense	-7	-24
cpr-4/R1	GC**GTCGAC**CTTCCAGAGAGATTGCTTTGA	cDNA and genomic	antisense	+170	+149
cpr-4/R2	TTGGAGCAGCAAGAAAGAACATC	cDNA and genomic	antisense	+483	+461
cpr-4/R2.5	CCGGTGCAGAATCCGCTC	genomic	antisense	+558	+541
cpr-4/R3	ACTGGTCCGTGGGCAATG	cDNA and genomic	antisense	+804	+787
cpr-4/R4	TATTATTATTCATATCTATATACACG	cDNA and genomic	antisense	+1116	+1091
cpr-4/5′FL	TTTCGGCGAGTACAGAGG	genomic	sense	see below	see below
cpr-4/3′FL	ATTATCTCTAACTTTATCG	genomic	antisense	see below	see below

Table iii. Custom primers for *cpr-5*

Name	Primer (Read 5′-3′)	Clones Sequenced With Primer	Sense of Primer	Position of 5′ End and 3′ End Bases of Each Primer With Respect to ATG Initiation Codon	
				5′ End	3’ End
cpr-5/1	GCTTTGGACCGCTGGACATCAAG	cDNA and genomic	sense	+123	+145
cpr-5/2	GGAGAAGATCACCAAGAAGCTGATGG	cDNA and genomic	sense	+156	+181
cpr-5/3	ACCGGAATGTTCAGCTGCG	cDNA and genomic	sense	+433	+451
cpr-5/4	CGAGGACTTCTACCAATACACC	cDNA and genomic	sense	+1026	+1047
cpr-5/5	CAGGGAGATTTCCATGTG	genomic	sense	+1303	+1320
cpr-5/R0.5	GAATAGCGGAGAGCTTCC	genomic	antisense	+22	+5
cpr-5/R1	GC**ACTAGT**CGAAGTGGTCTGGAATAGCGTC	cDNA and genomic	antisense	+260	+238
cpr-5/R2.5	GATTAAAACTACGGCGC	genomic	antisense	+563	+547
cpr-5/R2	GTTCACAGTCTCGCCGCATGGA	cDNA and genomic	antisense	+774	+753
cpr-5/R3	CCGTTGTCGACTCCCCAT	cDNA and genomic	antisense	+1127	+1110
cpr-5/5′FL	TTAGGTATTAGGCTCATC	genomic	sense	see below	see below
cpr-5/3′FL	GAAATCAAAGATTTGGG	genomic	antisense	see below	see below

Table iv. Custom primers for *cpr-6*

Name	Primer (Read 5′-3′)	Clones Sequenced With Primer	Sense of Primer	Position of 5′ End and 3′ End Bases of Each Primer With Respect to ATG Initiation Codon	
				5′ End	3′ End
cpr-6/0.5	CAAGCGACGACAACTTGC	genomic		-45	-29
cpr-6/1	GCATAGTGGTAGCAGCTTATTGCGC	cDNA and genomic	sense	+75	+99
cpr-6/2	GGACAAATATCGCAATCGTGAAATTGACTC	cDNA and genomic	sense	+127	+203
cpr-6/2.5	TGGGGATTGATGGGTGTC	genomic	sense	+368	+385
cpr-6/3	GATCCTGCTGGGCTTTCGG	cDNA and genomic	sense	+724	+742
cpr-6/4	CCAAAATGTGAAAAGAAGTGCG	cDNA and genomic	sense	+1077	+1098
cpr-6/5	AGTCTATGTTCACACCGG	cDNA	sense	+1262	+1326
cpr-6/5.5	CGGAGGAGGACACGCCG	genomic	sense	+1336	+1352
cpr-6/5.75	TCTGGAGTTGTTGGAGG	genomic	sense	+1481	+1497
cpr-6/6	CCGCCGCCACGTCTACG	genomic	sense	+1689	+1705
cpr-6/R0.5	TTGGCGTCGACGGTCAGC	genomic	antisense	2	-19
cpr-6/R1	GC**ACTAGT**CGTCTTTGTTTCTTAGCCGTCC	cDNA	antisense	+330	+266
cpr-6/R1.5	GCCTTGTCGTTCTCTCC	genomic	antisense	+363	+347
cpr-6/R1.75	AGCTTTCTGGAATGTCC	genomic	antisense	+657	+641
cpr-6/R2	GTTACATCCGAATCCACAGC	cDNA and genomic	antisense	+857	+838
cpr-6/R3	CCGTGAGTCATCAATTCT	cDNA and genomic	antisense	+1207	+1190
cpr-6/R4	ACACTCATCAACTCCTC	cDNA and genomic	antisense	+1471	+1455
cpr-6/R5	TAACTCATTCATTTTTATACATATAG	cDNA	antisense	+1788	+1763
cpr-6/5′FL	ATCGGTGTCATTCACTGA	genomic	sense	see below	see below
cpr-6/3′FL	TTCTAAATATATAGAATTG	genomic	antisense	see below	see below

Notes: The *cpr-3*/R1, *cpr-4*/R1, *cpr-5*/R1 and *cpr-6*/R1 primers all have restriction endonuclease sites engineered into their 5′ ends (shown in bold) because these primers were also used for the 5′ RACE experiments (see below). These sites did not therefore perform any function when these primers were used for sequencing.

The ~/5′FL and ~/3′FL primers for each of the four genes were used to sequence the second strand DNA of the 5′ and 3′ flanks of each of the four genes. The nucleotide sequences of these primers were derived from regions of readable sequence outwith the 5′ and 3′ flank sequence of each gene obtained from first strand DNA sequencing. This approach was used in order to obtain more nucleotide sequence confirmed in both strands for the 5′ and 3′ flanks of each gene without having to prepare more custom made primers. Since these primers were derived from readable sequence that was not adjacent to that obtained by first strand DNA sequencing of the 5′ and 3′ flanks of each gene, the positions of these primers with respect to the ATG initiation codon are not known accurately and are therefore not included in the above tables. However, the approximate positions of the ~/5′FL and ~/3′FL primers are upstream of the 5′ flank sequences and downstream of the 3′ flank sequences given for each gene, respectively (shown in Chapter 4, Figures 4.11, 4.12, 4.13 and 4.14). The primer *cpr-5*/5′FL is the notable exception to this rule. First strand sequencing of the 5′ flank of *cpr-5* revealed much repetitive sequence which negated the design of a specific primer for obtaining second strand DNA sequence. Accordingly, this primer was designed to unique sequence downstream of the region of repetitive sequence and within the region of DNA sequence obtained by first strand DNA sequencing. Therefore, the position of this primer with respect to the ATG initiation codon is known and is included in the appropriate table above.

2.11.3.3. Primers for sequencing the *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* constructs

During the construction of the *lacZ* gene fusions, a number of primers were used to check the sequence of the PCR amplified linkers and to check the final constructs. The names and sequences of these primers are listed below.

i) Primers used to sequence the PCR amplified linker regions

Name	Primer Sequence	Clones Sequenced
T3 and T7 17mers	sequence given above	*cpr-3*, *cpr-4* and *cpr-6* PCR amplified linkers in pBluescript II SK-
cpr-5 Primer 'X'*	5′-TTAGGTATTAGGCTCATC-3′	*cpr-5*::*lacZ*

SYNINT**	5′-GTTCTATGTTATGT TAGTATCATTCGAAACA-3′	*cpr-5*::*lacZ*

*This primer is also known as *cpr-5*/5′FL and is listed in the appropriate table above

**This primer anneals to the synthetic intron of the pPD21.28 *lacZ* fusion vector (see Chapter 5, Figure 5.21) and was used to obtain antisense sequence for the *cpr-5* PCR amplified linker.

ii) Custom made primers used to check *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* fusion constructs

Name	Primer Sequence
NUCLOC	5′-TCACCCACCGGTACCTTACGC-3′
M13 Reverse PCR Primer	5′-AACAGCTATGACCATGATTA-3′
SYNINT	sequence given above

Note: The orientation and relative positions of the NUCLOC, SYNINT and M13 Reverse PCR primers are shown schematically in Chapter 5, Figure 5.21.

2.12. 5′ Rapid amplification of cDNA ends (5′ RACE)

5′ RACE was performed with 2μg of total RNA isolated from mixed stage N2 Bristol *C.elegans* and 1pg of GibcoBRL control RNA using the GibcoBRL 5′ RACE kit (GibcoBRL) according to the manufacturer's protocol. The control RNA is an 891bp *in-vitro* transcribed region of the Chloramphenicol Acetyl Transferase gene, engineered to include a poly(A) tail. In brief, first strand cDNA was prepared from the total RNA (or from the control RNA) using pooled primers specific for each of the four genes (or the GSP1 control primer). The resulting cDNA was dC-tailed using Terminal Deoxynucleotidyl Transferase. The dC-tailed cDNA was subsequently amplified by nested PCR. The first round of the nested PCR used pooled primers specific to each of the four genes and the Anchor Primer supplied with the kit, which annealed to the dC-tailed region of the cDNAs. The cDNA generated from the GibcoBRL 5' RACE control

RNA was amplified in a separate reaction using the GSP2 control primer, the Universal Primer and the same cycle profiles as for the four genes.

Four independent PCR reactions were performed on the amplified products from the first round of nested PCR. Each reaction contained the Universal Primer, which annealed to the 5′ ends of the amplified products generated by the Anchor Primer in the first round of PCR, and a primer specific to one of the four genes. The products of the second round of nested PCR were electrophoresed through a 1.5% agarose gel and the predominant products from each reaction were gel purified. The gel purified products were cloned into pBluescript II SK- by means of *Spe*I or *Sal*I restriction endonuclease sites engineered into the 5′ ends of the universal and gene specific primers used in the final PCR amplification step. Thus the 5′ RACE products of *cpr-3* and *cpr-4* were cloned using *Sal*I sites and those of *cpr-5* and *cpr-6* were cloned using *Spe*I sites. The Anchor Primer, Universal Primer, control and gene specific primers used in the cDNA synthesis and amplification steps are listed below:

Name	Primers For cDNA Synthesis (GSP1)	Primers For 1st Round PCR (GSP2)	Primers For 2nd Round PCR (GSP3)
cpr-3	*cpr-3*/R3	*cpr-3*/R2	*cpr-3*/R1
cpr-4	*cpr-4*/R3	*cpr-4*/R2	*cpr-4*/R1
cpr-5	*cpr-5*/R3	*cpr-5*/R2	*cpr-5*/R1
cpr-6	*cpr-6*/R3	*cpr-6*/R2	*cpr-6*/R1

Note: The nucleotide sequences of these primers are listed in the appropriate table above (Section 2.11.3.2).

Control Primers: GSP1: 5′-TTGTAATTCATTAAGCATTCTGCC-3′
GSP2: 5′-CAUCAUCAUCAUGACATGGAAGCCATCACAGAC-3′

The Anchor Primer

*Spe*I

5′-CUACUACUACUAGGCCACG**CGTCGACTAGT**ACGGGIIGGGIIGGGIIG-3′

UDG cloning site *Sal*I anchor region

(Taken from the GibcoBRL 5′ RACE protocol)

The Universal Primer

5′-AGGCCACG**CGTCGACTAGT**ACGGG-3′

The Universal Primer used was synthesised in the laboratory because of limited quantities of the GibcoBRL Universal Primer supplied with the kit. The Universal Primer used is essentially identical to the one supplied with the kit but lacks the sequences required for UDG cloning since this was not required for the cloning approach employed.

2.12.1. Reaction conditions of the nested PCR amplification steps

2.12.1.1. First round of nested PCR

The reaction conditions used to amplify the dC-tailed cDNA in the first round of nested PCR are given below:

Reagent	Final Concentration
Taq polymerase 10x buffer (Promega Corporation)	1x
Taq polymerase (Promega Corporation)	1U/10µl
Acetylated BSA (Promega Corporation)	100µg/ml
$MgCl_2$	2.5mM
dNTPs (Promega Corporation.)	200µM
Anchor Primer	40nM
pooled gene specific primers for each of the four genes	100nM (each)

The reactions were made up in a final volume of 50µl using $sdiH_20$. Each reaction was overlaid with an approximately equal volume of mineral oil and PCR performed using an Omnigene thermocycler (Hybaid Ltd.) using the following conditions:

Step	Cycles	Temp(°C)	Time (min)
1	1	94	3
2.1	35	57	0.5
2.2		72	2
2.3		94	0.5
3.1	1	57	2
3.2		72	10

2.12.1.2. Second round of nested PCR

The reaction conditions used to amplify the products from the first round of the nested PCR are given below. 1/1,000th of the amplified products from the first round PCR were amplified using the following conditions:

Reagent	Final Concentration
Taq polymerase 10x buffer (Promega Corporation)	1x
Taq polymerase (Promega Corporation)	1 unit/10µl
Acetylated BSA (Promega Corporation)	100µg/ml
$MgCl_2$	2.5mM
dNTPs (Promega Corporation)	200µM
Universal Primer	300nM
gene specific primer for one of the four genes	100nM

The reactions were made up in a final volume of 50µl using $sdiH_20$. Each reaction was overlaid with an approximately equal volume of mineral oil and PCR performed using an Omnigene thermocycler (Hybaid Ltd.) using the same conditions as for the first round of PCR.

2.13. Semi-quantitative reverse transcription polymerase chain reaction (s-q rtPCR)

Semi-quantitative reverse transcription Polymerase Chain Reaction (sq-rtPCR) was used to analyse the mRNA abundance of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* through development. The 20 staged cDNA samples covering *C.elegans* larval and early adult development were supplied by Dr. I.L.Johnstone. These samples were generated in the following manner. Embryos were isolated from *C.elegans* and allowed to hatch in the absence of an *E.coli* OP50 food source to produce larvae arrested in L_1, the resulting arrested L_1 larvae were seeded onto NGM agar plates with *E.coli* OP50 as a food source and grown at 25°C. Samples of approximately 2000 worms were taken at 2 hour intervals and total RNA extracted. This yielded 20 RNA samples covering the 4 larval stages and the first 10 hours of adult development. Samples from older adults were not taken because they contained embryos. A fraction (1/100th) of the RNA from each time point was converted to first strand cDNA using Superscript II reverse transcriptase (GibcoBRL) and random hexamers. Mixed stage embryo RNA was isolated as described in Section 2.7.2. First strand cDNA was prepared from 2μg of the mixed stage embryo total RNA using 4μg of random hexamers (Pharmacia) and Superscript II reverse transcriptase enzyme and buffer (GibcoBRL) according to the instructions supplied with the enzyme.

The cDNA from each time point was amplified simultaneously using two sets of PCR primers, one set specific to the test gene (*cpr-3*, *cpr-4*, *cpr-5* or *cpr-6*) and the second set specific to the internal control transcript, *ama-1*. The primers used to amplify each transcript were designed either to distinguish cDNA from contaminating genomic DNA, or to prevent amplification of genomic DNA. Three pools of cDNA from each developmental stage were amplified with 23 cycles of PCR. The number of PCR cycles was chosen empirically to ensure that reactants were still in excess once the cycles were complete. The products were electrophoresed through 1.2%-2.0% agarose gels, Southern blotted and probed simultaneously with probes specific to the target and *ama-1* genes. Probes were initially generated by 5′ end-labelling the same primers used for the PCR amplification step (see Section 2.8.7.2). In later experiments, the probes were obtained by PCR amplification of the *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* cDNA clones, or by PCR amplification from pooled L_1 cDNA in the case of *ama-1*, using the same primer pairs as for the s-q rtPCR and gel purification of PCR products of the correct size (see below).

After autoradiography to detect signal, regions of the membrane corresponding to the amplified cDNA products of each gene were excised and radioactivity measured by counting in scintillant using a Beckman LS801 scintillation counter. The relative abundance of the test gene was then expressed as a ratio of the signal from the test gene to that of *ama-1* for each developmental stage. The mean and standard deviation of the ratios from each set of triplicate results were calculated.

2.13.1. PCR primers used for s-q rtPCR

The PCR primers used for the s-q rtPCR experiments for each of the four genes are described below:

Primer Pair	Primer Name	Primer Sequence
cpr-3 primer pair	*cpr-3*/1	listed above
	cpr-3/R3	listed above
cpr-4 primer pair	*cpr-4*/0.5	listed above
	cpr-4/R3	listed above
cpr-5 primer pair	*cpr-5*/1	listed above
	cpr-5/R3	listed above
cpr-6 primer pair	*cpr-6*/2	listed above
	cpr-6/R4	listed above
ama-1 Primer Pair	*ama-1*/1	5′-TTCCAAGCGCCGCTGCGCATTGTCTC-3′
	ama-1/2	5′-CAGAATTTCCAGCACTCGAGGAGCGGA-3′

The PCR primers used for the s-q rtPCR analysis of *cpr-4* proved to be non-specific and therefore a new PCR primer pair was designed. The names and nucleotide sequences of these primers are given below, in addition to the position of the 5′ end and 3′ end bases of these primers with respect to the ATG initiation codon of *cpr-4* (Chapter 4, Figure 5.12).

Primer Name	Primer Sequence	Primer Position	
		5′ Base	3′ Base
cpr-4/0.5PCR	5’-GCTCTATCTTGCTATTTGCTC TTTTACAAAAATGAAATAC-3′	-31	+58
cpr-4/R1.5PCR	5′-GGCTCCGTTGGAGGCGATGCAG-3′	+436	+415

2.13.2. The PCR reaction conditions used for s-q rtPCR

The same reaction conditions were used in the PCR amplification step of the s-q rtPCR experiments for each of the four genes. A general mix was prepared which contained the following:

Reagent	Final Concentration
Taq polymerase 10x buffer (Promega Corporation)	1x
Acetylated BSA (Promega Corporation)	100μg/ml
$MgCl_2$	1.5mM
dNTPs (Promega Corporation)	200μM
ama-1 primer pair	3ng/μl
gene specific primer pair	3ng/μl

For each gene, sufficient volumes of the above mix was prepared for all the required PCR reactions. The mix was heated to 70°C for 5min and then *Taq* DNA polymerase (Promega Corporation) was added to a final concentration of 1U/25μl. The reaction mixture was mixed and 24μl dispensed to a number of tubes each containing 1μl of the cDNA sample generated from each RNA time point, prewarmed to 80°C. Aerosol resistant tips were used for this to avoid cross contamination between samples. After the reaction mix had been added to each tube, and mixed with the template cDNA, an approximately equal volume of mineral oil was added to each tube to prevent evaporation and the PCR cycle profile activated immediately. In later experiments, the reactions were performed in a final volume of 50μl. The reaction conditions were identical and 48μl of reaction mix was added to 2μl of each cDNA sample, for each time point, giving the same ratio of template to reaction mix as for the 25μl reactions. In all cases the PCR amplification step of the s-q rtPCR experiments were performed using an Omnigene thermocycler (Hybaid Ltd.). The PCR cycle profile used for the s-q rtPCR experiments for *cpr-3*, *cpr-5* and *cpr-6* is listed below:

Step	Cycles	Temp(°C)	Time (min)
1	1	94	3
2.1	22	60	0.5
2.2		72	2
2.3		94	0.5
3.1	1	60	2
3.2		72	10

Note: For *cpr-4*, the temperature of the annealing steps, 2.1 and 3.1, were reduced from 60°C to 55°C.

2.13.3. PCR amplification of sequences required to probe the s-q rtPCR Southern blots

The same reaction conditions, primers and cycle profiles used for the PCR amplification step of s-q rtPCR were used to amplify the sequences required for probing the s-q rtPCR Southern Blots with the following alterations:

i) To obtain probes for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, approximately 200ng of the cloned cDNAs for these genes (in pBluescript II KS+) were used as template. For *ama-1*, 1µl of an unknown concentration of cDNA pooled from the 6 L_1 stage time points obtained from Dr I.L.Johnstone was used as template.

ii) The amount of *Taq* DNA polymerase used was increased to 1unit/10µl

iii) The number of cycles of steps 2.1.2.2 and 2.3 were increased from 22 to 35 cycles.

2.14. Colony PCR

Colony PCR was used to screen cloning steps performed during the generation of the *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* fusion constructs. The same reaction conditions and PCR cycle profiles were used in all cases and are listed below:

Reagents	Final Concentration
Taq polymerase 10x buffer (Promega Corporation)	1x
Taq DNA polymerase (Promega Corporation)	1 unit/25µl
Acetylated BSA (Promega Corporation)	100µg/ml
$MgCl_2$	1.5mM
dNTPs (Promega Corporation)	200µM
Primer A (listed below)	4ng/µl
Primer B (listed below)	4ng/µl

Table i. Primer A

Primer Name	Primer Sequence
M13 Reverse PCR Primer	listed above (Section 2.11.3.3)
cpr-3 Primer 'X' (*cpr-3*/5′FL)	listed above (Section 2.11.3.2, Table i)
cpr-4 Primer 'X'(*cpr-4*/5′FL)	listed above (Section 2.11.3.2, Table ii)
cpr-5 Primer 'X'(*cpr-5*/5′FL)	listed above (Section 2.11.3.2, Table iii)
cpr-6 Primer 'X'(*cpr-6*/5′FL)	listed above (Section 2.11.3.2, Table iv)

Table ii. Primer B

Primer Name	Primer Sequence
M13 Forward PCR Primer	5′-GTTGTAAAACGACGGCCAGT-3′
lacZ/2	5′-TCGCCATTCAGGCTGCGCAACTGTT-3′
SYNINT	listed above (Section 2.11.3.3)
NUCLOC	listed above (Section 2.11.3.3)

25µl of the reaction mixture was dispensed to as many 0.5ml eppendorf tubes as were required for the colony PCR screen. Isolated colonies from transformation plates or from streaked stock plates were picked using a wooden tooth pick. The pick was dipped in one of the eppendorf tubes, mixed and then dropped into 5ml of L-broth supplemented with the appropriate antibiotic to generate an overnight culture. The PCR reactions were overlaid with an approximately equal volume of mineral oil and PCR amplified using the following PCR cycle profiles:

Step	Cycles	Temp(°C)	Time (min)
1	1	94	3
2.1	20	55	0.5
2.2		72	3
2.3		94	0.5
3.1	1	55	3
3.2		72	15

Note: For colony PCR involving the large 5′ flank insert of *cpr-6*, the extension time of step 2.2 was increased to 4.5min. For colony PCR using the *cpr-5* Primer 'X', the annealing temperatures of steps 2.1 and 3.1 were reduced to 50°C.

2.15. High fidelity PCR

The 5′ flank linkers required for generation of the *lacZ* fusion constructs *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ*, were obtained using high fidelity PCR to reduce the probability of introducing PCR generated mutations. The principle differences to other PCR reactions were the use of a thermostable DNA polymerase with 3′-5′ exonuclease activity (proof-reading activity) and reduction of the total number of PCR cycles. The conditions used were as follows

Reagents	Final Concentration
Pwo DNA Polymerase (Boehringer Mannheim UK Ltd., Lewes, E.Sussex, UK)	1unit/40µl
Pwo DNA Polymerase 10x Buffer Including 20mM $MgSO_4$ (Boehringer Mannheim UK Ltd.)	1x
Dialysed BSA (Pharmacia)	100µg/ml
dNTPs (Promega Corporation)	200µM
Primer 'X' (listed below)	3ng/µl
Primer 'Y' (listed below)	3ng/µl
Template DNA (listed below)	1µg

Each reaction was made up in a final volume of 100µl using sdiH$_2$0 and overlaid with approximately 100µl of mineral oil. The PCR reactions were then amplified using the following PCR cycle profiles:

Step	Cycles	Temp (°C)	Time (min)
1	1	94	3
2.1	14	50	0.5
2.2		72	1
2.3		94	0.5
3.1	1	50	2
3.2		72	10

Nucleotide Sequences of Primer 'X' and 'Y' for Each Gene

Gene	Cloned Template	Primer 'X'	Primer 'Y'
cpr-3	CL#034/1.7/KS+	*cpr-3*/5′FL	5′-TAGA**GGATCC**GCCTAAATGTAAGAAAAATAATGG-3′
cpr-4	ZK1055/2.8/KS+	*cpr-4*/5′FL	5′-AGAA**CCCGGG**ATTTCTTCAAGACAATTACAATTATT-3′
cpr-5	W02B2/3/KS+	*cpr-5*/5′FL	5′-GAGA**CCCGGG**ACATTATGAGAGAAGTGTCTGCG-3′
cpr-6	C25B8/3.8/KS+	*cpr-6*/5′FL	5′-GAGC**GGATCC**TTCTAGAATTCAAATTTTCTCTAA-3′

Notes: The nucleotide sequences of Primer 'X' for each gene are given in the appropriate table of Section 2.11.3.2. The nucleotide sequences representing restriction endonuclease sites (*Bam*HI for *cpr-3* and *cpr-6, Xma*I for *cpr-4* and *cpr-5*) engineered into the 5′ end of Primer 'Y' for each gene are highlighted in bold.

2.16. PCR amplification of the 5′ ends of the cloned cDNAs of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The 5′ ends of the four cDNA clones were required for probing Southern blots during restriction endonuclease mapping of the genomic clones for each of the four genes, in order to determine the orientation of each gene within each clone. The 5′ ends of the cDNA clones were generated using PCR and the following conditions:

Reagents	Final Concentration
Taq polymerase 10x buffer (Promega Corporation)	1x
Taq DNA polymerase (Promega Corporation)	1 unit/10µl
Acetylated BSA (Promega Corporation)	100µg/ml
$MgCl_2$	2.5mM
dNTPs (Promega Corporation)	200µM
M13 Forward PCR Primer	2ng/µl
Gene Specific Primer (listed below)	2ng/µl
Template cDNA clone (listed below)	200ng

Template	Gene Specific Primer
cpr-3 cDNA clone (cm12b6)	*cpr-3*/R1 (Section 2.11.3.2, Table i)
cpr-4 cDNA clone (cm14e3)	*cpr-4*/R1 (Section 2.11.3.2, Table ii)
cpr-5 cDNA clone (cm04d10)	*cpr-5*/R1 (Section 2.11.3.2, Table iii)
cpr-5 cDNA clone (cm01a5)	*cpr-6*/R1 (Section 2.11.3.2, Table iv)

The reactions were made up in a final volume of 50µl and overlaid with an approximately equal volume of mineral oil prior to thermal cycling. The PCR cycle profiles used were as follows:

Step	Cycles	Temp(°C)	Time (min)
1	1	94	0.5
2.1	35	60	0.5
2.2		72	1
2.3		94	0.5
3.1	1	60	2
3.2		72	10

2.17. Generation of transgenic *C.elegans* lines and staining for ß-Galactosidase activity

2.17.1. Preparation of materials and worms for microinjection

Plasmid DNA for microinjection was extracted and purified using the Qiagen Plasmid Maxi or Plasmid Midi DNA kits (Qiagen Inc.) and was resuspended in TE. Each of the four *lacZ* fusion constructs, *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* fusion were diluted to 400µg/ml with sterile distilled water. The pRF4 marker plasmid, carrying the dominant *rol-6*(*su1006*) mutation was diluted to 200µg/ml in TE. Each of the diluted *lacZ* fusion constructs were mixed 1:1 with the diluted pRF4 plasmid to generate a final injection solution of 200µg/ml of *lacZ* fusion construct and 100µg/ml pRF4 marker plasmid.

Needles for microinjection were prepared from Aluminum Silicate glass capillaries (Clark Electromedical Instruments, Pangbourne, Reading, UK) which possessed glass filaments along their internal length to aid loading of the injection solution to the tip of the needle. Needles were pulled to the desired shape using a model 773 electrode puller (Campden Instruments Ltd., Sileby, Loughborough, UK). Immediately before use, the injection solution was centrifuged at 15,000rpm in a microcentrifuge (Heraeus Equipment Ltd.) and loaded into hand pulled borosilicate capillary tubes (Clark Electromedical Instruments) by drawing the injection solution up using a mouth pipettor. The DNA was then loaded into the needles by inserting the hand pulled capillary tube down the wide bore of the needle and depositing the injection solution just behind the needle tip, using gentle pressure with a mouth pipettor (see Section 2.17.2)

Worms to be injected were immobilised on dried agarose pads under paraffin oil essentially as described by A.Fire (Fire, 1986). The dried agarose pads were prepared on 60mm x 22mm glass coverslips (#1), by placing a single drop of a 3% agarose solution onto the coverslip, placing a 22mm x 22mm glass coverslip (#1) on the drop and applying gentle pressure. The smaller coverslip was removed and the agarose plug dried by baking at 80°C for 10min.

Young adult N2 Bristol *C.elegans* hermaphrodite worms were transferred to a partially desiccated NGM agar plate prior to injection. When the injection apparatus was ready, a sharpened platinum wire pick was used to transfer 5-15 of these worms to a dried agarose plug, overlaid with 2-3 drops of paraffin. The picked worms were allowed to touch the dried agarose plug and the pick was gently removed once a portion of the worm's body had adhered to the surface of the dried agarose. The immobilised worms were placed on the microscope stage and the injection procedure started immediately.

2.17.2. Microinjection procedure

The microinjection procedure and apparatus are essentially the same as that described by C.Mello *et al* (1991). Needles containing the injection solution were loaded into the microtool collar of a Narishige MO-202 Joystick Hydraulic Micromanipulator (Narishige Co Ltd., Tokyo 157, Japan) and were pressurised using an attached bottled nitrogen source. A pressure regulator and solenoid valve attached to the nitrogen cylinder were used to obtain the desired rate of flow of the injection solution. Pressure was activated by means of a foot operated switch linked to the solenoid valve.

Each large coverslip, carrying worms immobilised on the agarose plugs, was taped to the free sliding oil cushioned stage of a Zeiss Axiovert 100 inverted microscope fitted with Nomarski optics. Microinjections were performed at 400x magnification using a Zeiss Achroplan 40x air objective lens. Pressure was applied to the needle containing the injection solution using the foot operated switch. The micromanipulator was used to open the sealed needle by repeatedly touching the tip of the needle onto a piece of capillary tube placed on the agarose plug until a suitably high flow rate was observed, indicating that the needle possessed a sufficiently large bore to reduce blocking. Injections were performed by first focusing on the grainy cytoplasm in the centre of the distal gonad. The needle was then brought into the same focal plane and the microscope stage moved by hand to push the worm into the needle tip. Once the needle was touching the surface of the worm, the microscope stage was tapped and this usually resulted in the needle penetrating the cuticle and gonad. Nitrogen pressure was then applied to force the injection solution into the gonad immediately after penetration. The needle was rapidly withdrawn from the worm without reducing the nitrogen pressure, once the gonad had swollen visibly.

After injection, the worms were recovered by adding a drop of 1xM9 solution onto the agarose pad, causing the worms to float off the pad. The injected worms were then collected from the drop using a platinum wire pick and approximately 10 worms seeded onto a fresh NGM agar plate supporting an *E.coli* OP50 food source, by floating the worms onto a drop of sterile 1xM9 solution. The injected worms were incubated for 3-4 days at 20°C until F1 progeny with the *rol-6* phenotype were observed.

2.17.3. Isolation of transgenic lines

F1 progeny from the injected worms which exhibited the *rol-6* phenotype, conferred by the pRF4 marker plasmid, were seeded to fresh NGM agar plates supporting an *E.coli* OP50 food source (10-20 worms/plate). These plates were incubated for 3-4 days at 20°C until F2 progeny with the *rol-6* phenotype were observed. At least 3 F2 roller progeny were picked from each F1 plate that transmitted the *rol-6* phenotype to the F2 progeny and these worms were seeded onto fresh NGM agar plates supporting an *E.coli* OP50 food source (1 worm/plate). This process generated at least 3 populations of worms derived from a single animal from each original plate of transmitting F1 progeny. Of the populations generated from each F1 plate, the one with the greatest fraction of worms with the *rol-6* phenotype was selected as a source for the transmitting line. In this manner, at least 2 independent transmitting lines were generated for each injected *lacZ* fusion construct. These lines transmitted the roller phenotype to 10% - 90% of their offspring.

2.17.4. Fixing of transgenic worms

The worms were fixed essentially as described by J.M.Young and I.A.Hope (Young and Hope, 1993). 10 -20 worms with the roller phenotype were seeded onto large (9cm) NGM plates seeded with *E.coli* OP50 and incubated until the bacterial lawn was almost cleared (usually 4-5 days). The worms were washed off the plates using 1xM9 and pelleted at 5,000rpm for 10 seconds in a Heraeus microfuge. Approximately 100 worms in 10µl were placed on a subbed microscope slide (prepared by dipping the slide into a 0.01% poly-L-lysine solution and baking at 80°C for 10min) and overlaid with a 60mm x 22mm coverslip (#1). The slide was snap frozen on an aluminium block standing in dry ice and the cover slip removed with a scalpel blade. The slide was immediately immersed in methanol at -20°C and incubated at-20°C for 5min. The slide was transferred to acetone at -20°C and incubation continued at -20°C for a further 5min. The slides were air dried (worm side up) for 1 hour at room temperature using a large glass plate to act as a heat sink.

2.17.5. Staining of transgenic worms

Staining for ß-galactosidase activity was performed using the substrate X-gal essentially as described by A.Fire *et al* (1990). 40µl of sensitive staining solution (see below), prewarmed to 42°C, was added to the slide carrying the fixed worms, also prewarmed to 42°C, and overlaid with a 60mm x 22mm coverslip (#1). The coverslip was sealed to the slide using nail varnish and incubated at 42°C in a humidified environment until staining nuclei were visible (5min - 24 hours). The worms were examined at 100x, 200x and 400x magnification using a Zeiss Axioplan microscope and Zeiss Plan-Neofluar 10x/0.3, 20x/0.5 and 40x/0.75 air objectives, respectively. Photographs were taken using a Zeiss MC100 Spot camera and Kodak Ektachrome 64T colour reversal or Kodak Ektapress Gold II 100 colour negative 35mm film (Kodak Ltd.).

Sensitive Staining Solution

0.2M sodium phosphate buffer pH7.5
1mM $MgCl_2$
10mM redox buffer (50mM potassium ferricyanide, 50mM potassium ferrocyanide)
0.004% SDS
0.3% X-Gal (from a 30% (w/v) stock in dimethyl formamide, prepared fresh)

2.18. Computer programs

All DNA sequence computational data analyses were performed using programs from the GCG package (according to the Program Manual for the Wisconsin Package, version 8, (1994), Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA, 53711). All phylogenetic analyses were performed using programs from the PHYLIP package version 3.5c (Felsenstein, 1993) according to the program manuals supplied with the package. The ACEDB *C.elegans* database for Macintosh computers (MACACE) was used for all *C.elegans* genetic and physical map computer analyses.

Chapter 3

Chapter 3

Cloning Four Cathepsin B-like Genes from *C. elegans*

3.1. Introduction

3.1.1. The *C.elegans* physical map

In order to sequence a genome, it is necessary to generate an ordered physical map of clones covering that genome so that sequence data obtained from these clones can be accurately linked together. Furthermore, good alignment between such physical maps and their corresponding genetic maps are required in order to allow genes identified by sequencing to be accurately aligned to the genetic map.

Currently, the physical map of *C.elegans* comprises overlapping cosmid, lambda and yeast artificial chromosome (YAC) clones covering over 95 Mb (95%) of the genome. Initially, the physical map was assembled from cosmid and lambda clones using a restriction enzyme fingerprinting method to link overlapping clones into large sets known as "contigs" (Coulson *et al.,* 1986). This approach resulted in 90-95% of the *C.elegans* genome being cloned into 17,500 cosmids assembled into 700 contigs. The contigs could not be joined to each other because the linking clones required to do this were either not represented or severely under represented in the cosmid libraries. In order to overcome this problem, a YAC library of *C.elegans* genomic DNA was generated and a hybridisation strategy used to link the YAC and cosmid clones (Coulson *et al.,* 1988). Two sets of ordered grids were prepared; one of random YAC clones, representing approximately 2.5 genome equivalents, and the other of cosmids, which represented the contigs and unattached clones. Cosmids from the end of contigs were used to probe the YAC grid. YAC clones selected using this approach were isolated from host chromosomes by pulsed field gel electrophoresis and used to probe the cosmid grid. This proved to be a very effective approach for linking contigs together, with the number of contigs being reduced from 700 to 346 within seven months of the initiation of this project. This approach also facilitated the generation of a lower resolution physical map consisting of 958 ordered, overlapping YAC clones known as the YAC grid (Coulson *et al.,* 1991), since the hybridisation technique used also revealed the precise position and extent of the YAC clones with respect to the cosmid clones. At present, the majority of the physical map of *C.elegans* consists of a high resolution map of ordered cosmid and lambda clones covered by a lower resolution map of YAC clones represented on the YAC grid.

The physical and genetic maps of *C.elegans* are also well aligned, with over 90 Mb being positioned along the chromosomes using genetically mapped sequences and *in situ* hybridisation techniques (Sulston *et al.,* 1992). The accuracy of the alignment continues to increase as more genetically mapped sequences are determined.

3.1.2. The *C.elegans* genome sequencing project

The genome sequencing project started near the centre of chromosome III (Sulston *et al.,* 1992) because there was good cosmid coverage over several megabases of sequence and because the genetic map suggested that the central region of *C.elegans* chromosomes are gene rich. In the last report, sequences from 732 clones have been obtained, resulting in 22.5 Mb of finished sequence and an additional 15 Mb of sequence at various stages of completion (The *C.elegans* Genome Sequencing Consortium, pers. comm.). These sequences are primarily from the central regions of chromosomes II, III, IV and X. Once finished sequences have been obtained, the program GENEFINDER is used to find possible genes. These potential genes are then compared to sequences already in the public domain databases, such as GenBank, using the BLASTX program. Such analysis has revealed that approximately 46% of these predicted genes show significant similarities to sequences already in the databases. These predicted genes occur at an average density of one per 4.8 kb, for chromosomes II and III, or 6 kb, for the X chromosome (The *C.elegans* Genome Sequencing Consortium, pers. comm.). Based on these data, there are currently estimated to be approximately 13,526 (+/- 500) genes in *C.elegans* (The *C.elegans* Genome Sequencing Consortium, pers. comm.).

As a complementary approach to the *C.elegans* genome sequencing project, several groups are sequencing and mapping randomly selected cDNA clones from *C.elegans*. These projects generate partial sequence from cDNA clones which have been mapped to the YAC grid to produce sequence tags. These sequence tags can then be used to predict coding regions from the genomic sequence generated by the genome sequencing project. The initial results from two laboratories yielded approximately 1600 different cDNA sequences. McCombie *et al* (1992) isolated 585 random clones from a mixed stage cDNA library. These clones were partially sequenced from the 5′ and/or 3′ ends to produce 720 expressed sequence tags (ESTs). These ESTs were subsequently compared to sequences in GenBank. After redundancy was taken into consideration (where one gene is represented by more than one EST), 437 nuclear genes were identified. R.Waterston *et al* (1992) used a sorted cDNA library as a source of cDNA clones for random sequencing. The cDNA clones from this library were isolated in several cycles in which clones that hybridised to previously selected clones were discarded. This process generates a library that is partially normalised for expression levels, since isolation of multiple cDNA clones from the same gene will be selected against. For the same reason, this pre-selection screens out most highly homologous members of the same gene family. This screen has yielded sequences from 1,517 clones which appear to represent 1,194 individual genes. Comparison of the different cDNA sequences to sequences in the Genbank and EMBL databases revealed very similar results from each of these two screens. Each group identified 15 previously described *C.elegans* genes. Of the remaining sequences, both groups found that 25-30% of the cDNA sequences showed homology to previously isolated genes in the databases. In both screens, the majority of the clones obtained represent entirely novel sequences which will be very useful for determining coding sequence from the genome sequencing project. Both groups are currently mapping the partially sequenced cDNA clones to the *C.elegans* physical map (McCombie *et al.,* 1992; Waterston *et al.,* 1992). More recently, Yuji Kohara has implemented a project aimed at isolating all the cDNA species of *C.elegans* (Yuji Kohara, pers. comm.). In this project, the cDNA clones are not only being partially sequenced and physically mapped using the YAC grid, but also used to analyse expression patterns in whole mount embryos using a high throughput multi-well *in situ* hybridisation approach. Clearly, such projects will generate information that is not only useful for identifying genes from sequence generated by the genome sequencing project, but is also useful for helping to understand the biological roles of those genes identified.

3.1.3. Additional uses of the *C.elegans* physical map

The virtually complete physical map of *C.elegans* is not only an essential prerequisite for sequencing of the genome but is also an extremely useful tool in its own right. The physical map position of most cloned sequences can be determined rapidly by hybridisation to the YAC grid. This information may then be used to facilitate the rapid isolation of genomic clones corresponding to the sequences used to probe the YAC grid since cosmid and lambda clones can be selected which subtend the region covered by the hybridising YAC clones of the YAC grid, so avoiding the need to screen an entire cosmid or lambda library. The good alignment of the physical and genetic maps may also yield candidate mutant phenotypes for the genomic clone isolated. Furthermore, the ability to transform *C.elegans* allows the use of transgenic technology to determine whether an isolated genomic clone is capable of repairing such candidate mutant phenotypes. Thus the physical map of *C.elegans*, and its alignment to the genetic map, provides a potential route for rapidly cloning and mapping genes, some of which may be defined by mutations. This feature is especially useful since all the partially sequenced and mapped cDNA clones, generated in the screens discussed in Section 3.1.2, are being made available to researchers.

3.2. Results

3.2.1. Obtaining four cDNA clones encoding proteins with homology to cathepsin B

During their survey of expressed sequences from *C.elegans*, R.Waterston *et al* (1992) isolated and partially sequenced eight cDNA clones which showed significant homology to cathepsin B-like sequences in the public domain databases. A comparison of the partial sequence using the GAP program of the GCG package revealed that these cDNAs fell into four classes and were probably derived from four different genes. Between classes, the cDNA clones showed 37.8 - 62.6% identity. Within each class, the degree of identity rose dramatically to 91.6-100%. In the latter case, the lack of complete identity between some clones of the same class could be attributed to sequencing errors since all the clones were only sequenced on one strand. Indeed, the type of errors observed were consistent with this assumption, being primarily base insertions or deletions, resulting in frameshifts and disruption of the open reading frame for each of the clones, or a result of nucleotide residues whose base composition could not be determined, resulting in residues denoted as 'N'. Accordingly, four clones named cm12b6, cm14e3, cm04d10 and cm01a5, representing each of the four classes, were obtained from J.Sulston and A.Coulson for further analysis. Before any of the four clones were analysed further, the pRATII plasmids containing the cDNA inserts of these four clones were excised from the lambda SHLX2 vector and transfected into *E.coli* XL1-blue (see Chapter 2, Section 2.10). *E.coli* XL1-blue cells containing the plasmids were selected using the ampicillin resistance marker. The cDNA inserts from cm12b6, cm14e3, cm04d10 and cm01a5 in pRATII were subsequently subcloned into pBluescript II KS+ as *Apa*I/*Sac*I fragments for further characterisation.

The *Apa*I/*Sac*I cDNA inserts of the four selected cDNA clones, cm12b6, cm14e3, cm04d10 and cm01a5, were all used as probes to isolate genomic clones. Subsequent sequence analysis of the genomic clones (Chapter 4) revealed each to represent a distinct cathepsin B-like gene. These genes were named *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. A comparison of these genes to the cDNA sequence of cm12b6, cm14e3, cm04d10 and cm01a5 revealed complete base identity, interrupted by introns. Therefore the genes corresponding to the cDNA clones, cm12b6, cm14e3, cm04d10 and cm01a5, are subsequently referred to as *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively.

3.2.2. Confirming the physical map positions of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The results obtained from the physical mapping of a representative clone from each of the four classes of cathepsin B-like cDNA clones suggested that the gene, or genes, corresponding to each class of clone were located at distinct loci, with each cDNA clone hybridising strongly to a distinct region of the YAC grid (R.Waterston pers. comm.). However, several cDNA clones also hybridised weakly to additional regions of the grid. To confirm the preliminary YAC grid location of the four genes provided by R.Waterston, yeast strains carrying the YAC clones to which the four cDNA clones hybridised on the YAC grid were obtained from J.Sulston and A.Coulson for Southern blot analysis. Total genomic DNA was extracted from yeast strains carrying the YACs; Y69A3, Y7E11, Y48G5, Y50D4, Y55B10 and Y43B9. The restriction endonucleases *Bam*HI and *Eco*RI were used to digest 1μg of yeast genomic DNA, along with 5μg of wild type *C.elegans* Bristol N2 genomic DNA as a control. The digested DNA from each yeast strain was electrophoresed through a 0.8% agarose gel, with digested *C.elegans* genomic DNA in adjacent lanes. The DNA was Southern blotted to Hybond N (Amersham International plc), hybridised at medium stringency to the appropriate gel purified cDNA probe (radiolabeled by random priming) and washed twice at high stringency (Chapter 2, Section 2.8.8.3) prior to autoradiography.

According to the preliminary physical map data from R.Waterston; the genes *cpr-3* and *cpr-4*, corresponding to the cDNA clones cm12b6 and cm14e3, are contained within the YAC clones Y69A3 and Y7E11, respectively. The gene *cpr-5*, corresponding to the cDNA clone cm04d10, is contained within the YAC clones Y50D4 but also shows weak hybridisation to a second locus covered by Y48G5. The gene *cpr-6*, corresponding to the cDNA clone cm01a5, is contained within two overlapping YAC clones, Y43B9 and Y55B10. The results of each Southern blot analysis (Figures 3.1, 3.2, 3.3 and 3.4) indicate that these preliminary map positions are correct. Each cDNA clone produces strong signals and very similar hybridisation patterns when used to probe the control *C.elegans* genomic DNA and the yeast genomic DNA containing the appropriate YAC clone. Thus, the cDNA clones for *cpr-3*, *cpr-4* and *cpr-5* (cm12b6, cm14e3 and cm04d10) hybridise to Y69A3 (Figure 3.1), Y7E11 (Figure 3.2) and Y50D4 (Figure 3.3A), respectively, while the cDNA clone for *cpr-6* (cm01a5) hybridises to the two overlapping clones, Y55B10 and Y43B9 (Figure 3.4A and 3.4B). The cDNA clone for *cpr-5* (cm04d10) only produces weak signals when hybridised to Y48G5 (Figure 3.3B). The same pattern of weakly hybridising fragments is observed with *Eco*RI digested yeast DNA (lane 4) and control *C.elegans* DNA (lane 2), suggesting that a gene with some homology to the cm04d10 cDNA clone of *cpr-5* may be contained within Y48G5. However, no common pattern of hybridising fragments is observed for the *Bam*HI digested yeast genomic DNA (lane 3) and control *C.elegans* genomic DNA (lane 1) suggesting that this weak hybridisation is not reproducible with other digests. Together these data indicate that the *cpr-5* gene is not present at a second locus outwith the region covered by the Y50D4 YAC clone, but that there may be a gene in the region covered by the Y48G5 YAC clone which shows slight homology to the cm04d10 cDNA clone of *cpr-5*.

3.2.3. Analysis of the Southern blots of *C.elegans* genomic DNA and yeast DNA

The cDNA clones produce very similar, if not identical, hybridisation patterns when used to probe the digested control *C.elegans* genomic DNA (lanes 1 and 2, Figures 3.1, 3.2, 3.3 and 3.4) and yeast genomic DNA containing the appropriate YAC clones (lanes 3 and 4, Figures 3.1, 3.2, 3.3 and 3.4). Furthermore, the patterns of hybridisation observed for each of the four genes are in agreement with the patterns predicted from *Bam*HI and *Eco*RI restriction endonuclease sites revealed after sequencing the genomic clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. In addition to confirming the preliminary physical map data from R.Waterston, these data suggest that the genomic sequences corresponding to each of the cDNA clones are contained completely within the appropriate YAC clones. Furthermore, these data demonstrate the integrity of the genomic sequences corresponding to the four cDNA clones within each of the appropriate YAC clones. The hybridisation patterns observed for each gene will be discussed in more detail in the following sections.

3.2.3.1. The hybridisation pattern for *cpr-3*

The nucleotide sequence of *cpr-3* revealed the presence of a single *Bam*HI site towards the 5′ end of the coding region of this gene, but no *Eco*RI site. Thus, *Bam*HI digests of control *C.elegans* genomic DNA and yeast genomic DNA, containing Y69A3, are each expected to yield two hybridising bands, while *Eco*RI digests of DNA from the same two sources are each expected to yield a single hybridising band. Initially, the results obtained experimentally do not appear to be in total agreement with the expected results (Figure 3.1) since the *Bam*HI digested *C.elegans* genomic DNA (lane 1) generates only a single hybridising band of approximately 23 kb, rather than the two bands expected. However, a closer analysis of the autoradiograph reveals that the perceived single hybridising band in lane 1 actually comprises two bands migrating very close together. Furthermore, the difference seen between the signal intensities of these two bands in lane 1 is also observed for the yeast DNA in lane 3. Therefore the only discrepancy in the results is due to a reduction in size of one of the Y69A3 hybridising bands from approximately 23 kb to 10 kb. These data suggest that this fragment, and therefore the *cpr-3* gene, is located at one end of the cloned insert of Y69A3. If some of this fragment was replaced by YAC vector sequence, as a result of the enzymes used to clone the insert, then a proximal *Bam*HI site in the YAC vector arm could generate the observed decrease in size. The *Eco*RI digests, in lanes 2 and 4, both produced an identical and much smaller single hybridising band of approximately 4 kb. This suggests that the sequences flanking the gene are still present in the Y69A3 clone, since any replacement with vector sequence must have occurred outwith the *Eco*RI sites.

3.2.3.2. The hybridisation pattern for *cpr-4*

The nucleotide sequence of *cpr-4* revealed the presence of a single *Bam*HI site near the centre of the coding region of this gene, and two *Eco*RI sites 670 bp apart. Thus, *Bam*HI digests of control *C.elegans* genomic DNA and yeast genomic DNA, containing Y7E11, are each expected to yield two hybridising bands. *Eco*RI digests of DNA from the same two sources are each expected to yield three hybridising bands one of which should be 670 bp in size. The hybridisation pattern predicted from the nucleotide sequence of *cpr-4* is also obtained experimentally (Figure 3.2). However, there is a slight discrepancy in the size and hybridisation intensity of the largest band generated by *Bam*HI digestion of control *C.elegans* DNA (lane 1) and yeast genomic DNA containing Y7E11 (lane 3). Such a discrepancy may be genuine, reflecting the presence of intervening YAC vector sequences, as was observed for *cpr-3*, or it may be an artefact, since bands of high molecular weight DNA do occasionally migrate aberrantly in digests of genomic DNA and this may be caused by differences in the salt or DNA concentrations between samples. The largest hybridising fragment generated by *Bam*HI digested yeast DNA containing Y7E11(lane 3) is also fainter than that generated by *Bam*HI digested control *C.elegans* genomic DNA (lane 1) and this is probably a result of inefficient transfer of the larger DNA fragment to the Hybond N filter during Southern blotting.

3.2.3.3. The hybridisation pattern for *cpr-5*

The nucleotide sequence of *cpr-5* revealed the presence of a single *Bam*HI site near the centre of the coding region of this gene, and one *Eco*RI site towards the 5′ end of the coding region. Thus, *Bam*HI and *Eco*RI digests of both control *C.elegans* genomic DNA and yeast genomic DNA, containing Y50D4, are all expected to yield two hybridising bands of unknown size and this hybridisation pattern is obtained experimentally (Figure 3.3A). However, the signal from the 6.6 kb hybridising band of *Eco*RI digested yeast genomic DNA containing Y50D4 (lane 4) is somewhat less than that of the corresponding band from *Eco*RI digested control *C.elegans* genomic DNA (lane 2). This variation may be a result of inefficient transfer of the Y50D4 6.6 kb hybridising fragment to the Hybond N filter during Southern blotting, caused by local variations in transfer efficiency across the surface of the filter.

As well as the expected hybridising bands, strong signals from additional hybridising bands are also obtained with the *Bam*HI digested control *C.elegans* genomic DNA (lane 1). These additional bands are probably due to slight partial digestion since comparable signals are not obtained with *Eco*RI digested control *C.elegans* genomic DNA (lane 2) and are therefore unlikely to be a result of cross hybridisation to other sequences within the *C.elegans* genome. It should be noted that the same additional hybridising bands are also observed for the *Bam*HI digested *C.elegans* genomic DNA used as a control for the Southern blot of yeast DNA containing the Y48G5 YAC clone (lane 1, Figure 3.3B). This is not in disagreement with the conclusion that partial digestion may have caused the presence of these additional bands since the *Bam*HI digested *C.elegans* genomic DNA controls used for the Southern blots of Y50D4 and Y48G5 were obtained from the same digest.

3.2.3.4. The hybridisation pattern for *cpr-6*

The nucleotide sequence of *cpr-6* revealed the presence of a single *Bam*HI site near the centre of the coding region of this gene, and three *Eco*RI sites. Thus, *Bam*HI digests of control *C.elegans* genomic DNA and yeast genomic DNA, containing Y55B10 or Y43B9, are each expected to yield two hybridising bands, while *Eco*RI digests of

DNA from the same two sources are each expected to yield four hybridising bands, two of which are expected to be 380 bp and 1060 bp in size. This predicted hybridisation pattern is obtained experimentally for both Y55B10 and Y43B9 (Figures 3.4A and 4.4B) suggesting that the genomic sequences of *cpr-5* corresponding to the cm01a5 cDNA clone are completely contained within both these YAC clones.

3.2.4. The copy number of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The Southern blots of *Bam*HI and *Eco*RI digested *C.elegans* genomic DNA and yeast genomic DNA containing the appropriate YAC clones, probed with the cm12b6, cm14e3, cm04d10 or cm01a5 cDNA clones, indicate that the four cDNA clones hybridise strongly to four distinct loci (Figures 3.1, 3.2, 3.3 and 3.4). Furthermore, the Southern blot analyses produced simple hybridisation patterns for all four genes which could be completely explained by the *Bam*HI and *Eco*RI restriction endonuclease sites revealed during later sequencing of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, corresponding to the cDNA clones cm12b6, cm14e3 , cm04d10 and cm01a5, respectively. These data strongly suggest that each of these four genes are present as single copies in the genome of *C.elegans* at four distinct loci. From these data alone, it is not possible to discount each of these genes being tandemly repeated at each of their loci. However, later restriction endonuclease mapping of subclones containing each of the four genes strongly suggest this is not the case.

3.2.5. *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* do not cross hybridise significantly

As discussed in Section 3.2.2, the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 all generate the greatest signal intensities when hybridising to sequences corresponding to their genomic counterparts, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively. These data suggest that the nucleotide sequences of the four genes are sufficiently diverged to prevent significant levels of cross hybridisation of their cDNA clones to each other, or to other cathepsin B-like sequences in the *C.elegans* genome. However, the cDNA clones of *cpr-5* and *cpr-6* (Figures 3.3 and 3.4), do also generate some additional faint signals caused by cross hybridisation to other sequences in the *C.elegans* genome.

This suggests that there may be additional members of the cathepsin B-like multigene family within the genome of *C.elegans*. Alternatively, the additional faint signals may be the result of low levels of cross hybridisation between each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. However, this is unlikely to be the case since later sequence analysis suggest that the nucleotide sequences of these four genes are sufficiently diverged to make cross hybridisation to one another very unlikely (Chapter 4, Section 4.2.5).

The above data also indicate that the hybridisation and wash stringencies used were sufficiently high to distinguish between each of the four genes. These stringencies were therefore used in the experiments described in the subsequent sections of this chapter, in order to isolate genomic clones containing the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*.

3.2.6. Selecting cosmid clones for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

In addition to the overlapping YACs, the *C.elegans* physical map is represented by a set of overlapping cosmid (and lambda) clones with positional correlation between the YAC and cosmid sets (see introduction to this chapter). In order to obtain cloned genomic sequences of each of the four genes, sets of overlapping cosmids were selected on the basis of the positional information gained from the YAC grid physical map data (R.Waterston pers. comm.) and Southern blot data (Figure 3.1, 3.2, 3.3 and 3.4). These clones were then obtained from J.Sulston and A.Coulson at the Sanger Centre.

The physical map data indicate that the *cpr-3* cDNA clone (cm12b6) hybridises to two different YACs, Y46B6 (Figure 3.5A) and Y69A3 (Figure 3.5B), at distinct loci. However, the region covered by Y46B6 is completely overlapped by two additional YAC clones, Y43F3 and Y46C2. Therefore, hybridisation of cm12b6 to Y46B6 should be accompanied by hybridisation to one of these two YAC clones (Figure 3.5A). No such hybridisation was reported and therefore the hybridisation of cm12b6 with Y46B6 on the YAC grid must be artefactual. The Southern blot data obtained using the *cpr-3* cDNA clone (cm12b6) as a probe (Figure 3.1) agreed with this since these data suggest that *cpr-3* is single copy and is contained within the cloned insert of Y69A3. Some regions of the Y69A3 YAC clone are also covered by two additional overlapping YAC clones, Y59A11 and Y45G4 (Figure 3.5B), yet the physical map data suggested that cm12b6 only hybridised to Y69A3. This suggests that the *cpr-3* gene is contained within the region of the Y69A3 clone not covered by Y59A11 and Y45G4. The four cosmids ZK1037, C29F3, T09D4 and R14F11, which cover this region, were therefore selected and requested from J.Sulston and A.Coulson at the Sanger Centre. However, I was unable to obtain R14F11 because the Sanger Centre experienced problems growing it. Furthermore, there is a region of the Y69A3 YAC clone that is not covered by cosmid clones (indicated in Figure 3.5B). Consequently, I was unable to obtain cosmid clones covering the entire region of Y69A3 expected to contain the *cpr-3* gene.

The physical map data indicated that the *cpr-4* cDNA (cm14e3) hybridises to two overlapping YAC clones, Y7E11 and Y6G12 (Figure 3.6). The Southern blot data obtained using the *cpr-4* cDNA clone (cm14e3) as a probe (Figure 3.2) agree with this since these data suggest that *cpr-4* is single copy and is contained within the cloned insert of Y7E11. Accordingly, the cosmid clones, C40C4, ZK1055, F44C4 and T10H9, which cover the region of overlap of Y7E11 and Y6G12 (Figure 3.6), were selected and requested from J.Sulston and A.Coulson at the Sanger Centre.

The physical map data supplied with cm04d10 were not obtained using this clone but rather the cDNA clone cm5d10. A comparison of the partial sequence data of these two clones reveals that they share 98.7% identity and are therefore assumed to be derived from the same gene, for the reasons discussed in Section 3.2.1. The physical map data indicated that cm5d10 hybridises to two different regions of the YAC grid. This cDNA clone produced a strong hybridising signal within the region covered by Y50D4 (Figure 3.7A) and a weak hybridising signal within the region covered by Y48G5 (Figure 3.7B). The Southern blot analysis obtained using the *cpr-5* cDNA clone (cm04d10) as a probe suggests that *cpr-5* is single copy and contained within the cloned insert of Y50D4 (Figure 3.3A). However, a weak hybridising band was also observed when Southern blots of *Eco*RI digested *C.elegans* genomic DNA and yeast DNA containing the Y48G5 YAC clone were probed with cm04d10 (Figure 3.3B). Together, the physical map and Southern blot data suggest that there may be a gene with some homology to the cm5d10 and cm04d10 cDNA clones in the region covered by the Y48G5 YAC clone. The YAC grid data indicates that cm5d10 hybridises strongly to three overlapping YAC clones, Y50D4, Y50A6 and Y54C3 (Figure 3.7A). Therefore, three cosmids that cover this overlap, W02B2, T21H3 and W06C3, were requested from

J.Sulston and A.Coulson at the Sanger Centre. The YAC grid data also indicates that cm5d10 hybridises weakly to the three overlapping YAC clones Y48G5, Y19H3 and Y41A2 (Figure 3.7B). Therefore, three cosmids that cover this overlap, F09F2, AD10 and F58E8, were also requested from J.Sulston and A.Coulson at the Sanger Centre.

The physical map data indicate that the *cpr-6* cDNA clone (cm01a5) hybridises to four overlapping YAC clones; Y43B9, Y55B10, Y44C4 and Y42G7 (Figure 3.8). The Southern blot data obtained using the *cpr-6* cDNA clone (cm01a5) as a probe (Figure 3.4A and 3.4B) agreed with this since these data suggest that *cpr-6* is single copy and is contained within the cloned inserts of both Y43B9 and K11B10. Therefore, two cosmids, C25B8 and K11B10, covering this region of overlap, were selected and requested from J.Sulston and A.Coulson at the Sanger Centre.

3.2.7. Isolating genomic clones containing *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The selected cosmid clones (Section 3.2.6) were screened by colony hybridisation (Chapter 2, Section 2.8.8.6) using the appropriate cDNA clones as probes. For *cpr-3*, the results indicated that none of the selected cosmid clones hybridised to the *cpr-3* cDNA clone (cm12b6). For *cpr-4*, the two overlapping cosmid clones, ZK1055 and F44C4, were found to hybridise to the *cpr-4* cDNA clone (cm14e3). For *cpr-5*, the W02B2 cosmid clone was found to hybridise to the *cpr-5* cDNA clone (cm04d10). The AD10 cosmid also produced a very weak signal with cm04d10, suggesting that there might be a related gene present on this cosmid, subtending the region covered by the YAC Y48G5, as was suggested from the physical map data and the Southern Blot analysis of digested yeast genomic DNA containing Y48G5. For *cpr-6*, the overlapping cosmid clones, C25B8 and K11B10, were found to hybridise to the *cpr-6* cDNA clone (cm01a5). Thus, the cosmid clones; ZK1055 and F44C4 for *cpr-4*, W02B2 and AD10 for *cpr-5* and C25B8 and K11B10 for *cpr-6* were selected for further analysis.

From the colony hybridisation, it was apparent that *cpr-3* was not contained within the cosmid clones chosen. Therefore, I screened a *C.elegans* Bristol N2 genomic DNA lambda EMBL4 library obtained from I.Johnstone. Serial dilutions were used to assay the titre of the library, which was found to be 300,000 pfu/ml (Chapter 2, Section

2.9.1). The number of plaques required for the library screen was determined using the formula:

$$N = \frac{ln\,(1-P)}{ln\,(1-(1/n))} \qquad \text{(Clark and Carbon, 1976)}$$

Where N = the number of recombinants, P = the probability of including any DNA sequence in a random library of N independent recombinants, n = the size of the genome relative to a single cloned fragment.

Therefore, assuming: $P = 0.99$, $n = \frac{1 \text{ X } 10^8}{1.5 \text{ X } 10^4}$ $\frac{\text{(\textit{C.elegans} genome size)}}{\text{(average insert size)}}$

$N = 30{,}699$ recombinants

To ensure an even greater probability that the *cpr-3* gene corresponding to cm12b6 was included, 90,000 plaques were used in the primary screen. This yielded eight putative positive clones. These eight clones were selected and subjected to two further rounds of plaque purification. Seven positives clones named CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314 remained after plaque purification and these clones were selected for further analysis as genomic clones of *cpr-3*.

3.2.8. Isolation of subclones from the positive cosmid and lambda clones

I decided to subclone the genes from the cosmid and lambda clones into pBluescript II KS- for restriction endonuclease mapping. This approach had the advantage in that the more complex approaches required for mapping cosmid and lambda clones could be avoided. Large fragments which hybridised to the appropriate cDNA clones were subcloned into pBluescript II KS- to reduce the danger of losing sections of the genes, or their flanking sequences, during this subcloning step.

3.2.8.1. Obtaining suitable fragments for subcloning

DNA from the six selected cosmid clones was extracted using the alkaline hydrolysis method and purified by caesium density gradient centrifugation (Chapter 2, Section 2.6.4). The six cosmid clones were digested with a variety of restriction endonucleases, whose recognition sequences were also present in the pBluescript II KS- polylinker, and electrophoresed through 0.8% agarose gels. The gels were Southern blotted, probed with the appropriate cDNA (radiolabeled by random priming) and washed at high stringency prior to exposure to X-ray film.

For *cpr-4* (Figure 3.9), restriction endonuclease digests of ZK1055 DNA and F44C4 DNA probed with the *cpr-4* cDNA clone (cm14e3) generated almost identical hybridisation patterns with the same enzyme. Most of the differences observed were due to partial digestion, confirmed by viewing the ethidium bromide stained agarose gels. Partial digests were observed for; ZK1055 DNA digested with *Xba*I (lane 2, Figure 3.9A), ZK1055 DNA digested with *Sac*II (lane 6, Figure 3.9A), ZK1055 DNA digested with *Spe*I (lane 9, Figure 3.9A) and both ZK1055 and F44C4 DNA digested with *Kpn*I (lane 10, Figures 3.9A and 3.9B). The overall similarity in the hybridisation patterns, generated by probing the digested ZK1055 and F44C4 cosmid DNA, suggest that each cosmid carries a copy of the *cpr-4* gene, and that the gene is contained entirely within each of the cosmids, assuming that the *cpr-4* cDNA clone (cm14e3) used as a probe is nearly full length. However, the partial digest of ZK1055 cosmid DNA with *Spe*I (lane 9, Figure 3.9A) generates three hybridising bands, as opposed to the single hybridising band resulting from complete digestion of F44C4 cosmid DNA by *Spe*I (lane 9, Figure 3.9B). The faintest of these three hybridising bands does not appear to correspond to a partial digestion product, suggesting that there may be additional *cpr-4* sequences within ZK1055 not present in F44C4. This seems very unlikely since the *Sac*I digests (lane 3, Figures 3.9A and 3.9B) and *Pst*I digests (lanes 5, Figures 3.9A and 3.9B) of ZK1055 and F44C4 both produced single hybridising bands of approximately 4.5 kb and 3.5 kb respectively, for both cosmids, suggesting that *cpr-4* was flanked by *Sac*I and *Pst*I sites proximal to this gene. One would not expect such sites to be conserved between the two cosmids if extra sequences from the *cpr-4* gene were present in ZK1055 but absent in F44C4. This observation suggests that the anomaly observed was probably an artefact of partial digestion of ZK1055 by *Spe*I. However, I decided to use ZK1055 as a source of DNA to isolate subclones containing the *cpr-3* gene, assuming that this cosmid clone might possess additional sequences from this gene.

For *cpr-5* (Figures 3.10A and 10B), only restriction endonuclease digests of W02B2 DNA generated a hybridisation pattern when probed with the *cpr-5* cDNA clone (cm04d10). This suggests that the weak hybridisation of the *cpr-5* cDNA clone (cm04d10) to AD10 observed during colony hybridisation (Section 3.2.7) was probably an artefact, since the wash stringency was the same as that used when probing Southern blots of *Eco*RI digested *C.elegans* genomic DNA and yeast genomic DNA containing Y48G5, which both yielded a weak hybridising band (Figure 3.3B)

For *cpr-6* (Figure 3.11A and 3.11B), only restriction endonuclease digests of C25B8 DNA generated a hybridisation pattern when probed with the *cpr-6* cDNA clone (cm01a5). This indicates that the hybridisation of the *cpr-6* cDNA clone (cm01a5) to K11B10 observed during colony hybridisation (Section 3.2.7) was probably an artefact. Analysis of the ethidium bromide stained agarose gels prior to Southern blotting revealed that partial digestion of C25B8 cosmid DNA had occurred with *Xba*I (lane 2, Figure 3.11A) and *Sal*I (lane 4, Figure 3.11A).

The above digests indicate that *cpr-4* is contained within ZK1055 and F44C4, *cpr-5* is contained within W02B2 and *cpr-6* is contained within C25B8. For all three genes, simple hybridisation patterns are obtained, suggesting that these genes are all single copy within the appropriate hybridising clones. For *cpr-4*, I decided to subclone the 6.5 kb hybridising fragment produced by *Hin*dIII digestion of ZK1055 (lane 8, Figure 3.9A). However, for *cpr-5* and *cpr-6*, none of the single digests of W02B2 and C25B8 generated single hybridising fragments of a useful size for subcloning, these fragments being either too large for efficient subcloning into pBluescript II KS- or too small to aid subcloning of the complete genes including their flanking sequences.

For *cpr-5* and *cpr-6*, double digests were performed using a variety of restriction enzymes which previously produced large (greater than 12 kb) hybridising fragments with single digests of the cosmids W02B2 and C25B8. Double digests were performed using *Apa*I, *Sac*I and *Pst*I. The digestion products were electrophoresed through 0.7% agarose gels, Southern blotted and probed with the appropriate cDNA as previously described. The results demonstrate that *Pst*I/*Sac*I double digests of both W02B2 DNA and C25B8 DNA produce hybridising fragments of approximately 10 kb useful for subcloning *cpr-5* and *cpr-6* respectively.

The agarose gels used to identify hybridising fragments suitable for subcloning *cpr-4*, *cpr-5* and *cpr-6* from ZK1055, W02B2 and C25B8 could not accurately resolve the size of these fragments. Therefore the digests that were to be used for subcloning DNA fragments containing *cpr-4*, *cpr-5* and *cpr-6* from ZK1055, W02B2 and C25B8 were repeated and electrophoresed a greater distance through a 0.7% agarose gel. The agarose gel was Southern blotted, probed with the appropriate cDNA clone and washed at high stringency prior to autoradiography. The results (Figure 3.12) confirm that a 6.8 kb *Hin*dIII fragment from cosmid ZK1055 contains the *cpr-4* gene, an 11 kb *Pst*I/*Sac*I fragment from cosmid W02B2 contains the *cpr-5* gene and a 10.7 kb *Pst*I/*Sac*I fragment from cosmid C25B8 contains the *cpr-6* gene.

3.2.8.2. Subcloning fragments containing *cpr-4*, *cpr-5* and *cpr-6*

A 'shotgun' approach was used to subclone the 6.8 kb *Hin*dIII fragment from cosmid ZK1055. the 11 kb *Pst*I/*Sac*I fragment from cosmid W02B2 and the 10.7 kb *Pst*I/*Sac*I fragment from cosmid C25B8. With this approach, the cosmid DNA was digested with the appropriate enzyme(s) and the whole digest used in a ligation reaction with pBluescript II KS- which had been digested with the appropriate enzyme(s) and gel purified. Transformants were then isolated, using ampicillin resistance and blue/white colony selection. Colony hybridisation (Chapter 2, Section 2.8.8.6) was used to screen 50 - 100 transformants for the appropriate insert using the four cDNA clones. The colony hybridisation screen yielded; nine potential positive subclones containing the 6.8 kb *Hin*dIII fragment from ZK1055, seven potential positive subclones containing the 11 kb *Pst*I/*Sac*I fragment from W02B2 and-three-potential positive subclones containing the 10.7 kb *Pst*I/*Sac*I fragment from C25B8.

Plasmid DNA was prepared from overnight cultures of the potential positive subclones. The plasmid DNA was digested with the same restriction endonucleases originally used to prepare the ZK1055, W02B2 and C25B8 DNA for subcloning. The digestion products were electrophoresed through 0.7% agarose gels and were visualised by ethidium bromide staining. These digests revealed that five of the nine potential positive subclones from ZK1055 and one of the three potential positive subclones from C25B8 contained single inserts of approximately the correct size. However, while six of the seven potential positive subclones from W02B2 contained an insert of the correct size, additional inserts were also seen in these six cases. Therefore, I decided to subclone the 11 kb *Pst*I/*Sac*I fragment from W02B2 using a more standard approach. One of the six subclones obtained from the colony hybridisation screen was selected as a convenient source of the 11 kb *Pst*I/*Sac*I fragment because this insert could be easily isolated from the other contaminating inserts. The 11 kb insert was released from the vector by digestion with *Pst*I and *Sac*I restriction endonucleases, gel purified and ligated into pBluescript II KS-. Ten colonies were selected after transformation and overnight cultures prepared. The plasmid DNA was extracted, digested with *Pst*I and *Sac*I restriction endonucleases, electrophoresed through a 0.7% agarose gel, Southern blotted and probed using the *cpr-5* cDNA clone (cm04d10). Analysis of the ethidium bromide stained gel prior to Southern blotting and analysis of the autoradiograph generated after probing revealed that seven of the 10 clones screened carry the desired 11 kb insert from W02B2, and no other contaminating inserts.

Plasmid DNA from the five selected potential positive subclones from ZK1055, and the single potential positive subclone from C25B8, were digested with the appropriate restriction endonucleases. The digests were electrophoresed through a 0.7% agarose gel, Southern blotted and probed with the appropriate cDNA clone. The results from this screen demonstrated that four of the five subclones from ZK1055 possess a single 6.8 kb *Hin*dIII fragment which hybridised to the *cpr-4* cDNA clone (cm14e3) and thatonesubclone from C25B8 possesses a single 10.7 kb *Pst*I/*Sac*I insert which hybridised to the *cpr-6* cDNA clone (cm01a5).

In summary, the genes *cpr-4*, *cpr-5* and *cpr-6*, corresponding to the cDNA clones cm14e3, cm04d10 and cm01a5 respectively, were subcloned into pBluescript II KS- as large (6.8 - 11 kb) fragments. One clone for each gene was selected for further analysis and these were named ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, corresponding to the cDNA clones cm14e3, cm04d10 and cm01a5 respectively. Each of the selected subclones contains the gene to which the appropriate cDNA clone hybridises. Digestion with the appropriate restriction endonuclease enzymes and agarose gel electrophoresis indicate that none of these subclones contain additional, non-contiguous inserts.

3.2.8.3. Isolating the *cpr-3* gene from lambda clones

DNA from the seven positive lambda clones CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314, obtained from the *C.elegans* Bristol N2 genomic DNA lambda EMBL4 library screen (Section 3.2.7), was digested with the *Eco*RI and *Bam*HI restriction endonucleases. The digested DNA was electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N (Amersham International plc) and probed with the *cpr-3* cDNA clone (cm12b6).

Figure 3.13A shows the DNA fragments produced after agarose gel electrophoresis of the seven lambda clones digested with *Eco*RI. *Eco*RI digestion of all seven lambda clones generates two fragments of 20 kb and 9 kb, representing the left and right arms of the vector respectively. Analysis of the other fragments generated by *Eco*RI digestion reveals that all seven lambda clones share common bands of approximately 3 kb and 4 kb, suggesting that the inserts of the seven lambda clones all overlap one another. Additional bands unique to individual clones, or groups of clones, were also seen and at least three groups of lambda clones could be distinguished on the basis of larger, easily visible bands. The three groups are represented by CL#012, CL#034 and CL#056. The CL#012 clone appears to be unique, CL#034 may be identical to CL#078 and CL#0910. CL#056 may be identical to CL#01112 and CL#01314. However, smaller bands were obscured by contaminating *E.coli* genomic DNA making it impossible to determine whether the clones within the latter two groups were truly identical to each other. Figure 3.13B shows the DNA fragments produced after agarose gel electrophoresis of the seven lambda clones digested with *Bam*HI. These results reveal that CL#034 and CL#0910 share a common 2 kb band absent in the clone CL#078. Thus the group of lambda clones comprising CL#034, CL#078 and CL#0910, determined from the *Eco*RI digests, may be divided further into two groups. Thus the seven lambda clones can be organised into at least four groups based on the pattern of DNA fragments generated by *Eco*RI and *Bam*HI digests. The autoradiographs obtained after probing the Southern blots of these agarose gels with the *cpr-3* cDNA clone (cm12b6) are shown in Figures 3.14A and 3.14B. The autoradiograph obtained from *Eco*RI digested DNA from the seven lambda clones (Figure 3.14A) reveals that all these clones contain a common 4 kb hybridising fragment. A 4 kb hybridising fragment was also observed when the *cpr-3* cDNA clone (cm12b6) was used to probe Southern blots of *Eco*RI digested *C.elegans* Bristol N2 genomic DNA and yeast DNA containing the Y69A3 YAC clone (Figure 3.1). The autoradiographic data obtained from *Bam*HI digested DNA from the seven lambda clones (Figure 3.14B) reveals that CL#034 and CL#0910 share a common 2 kb hybridising band not present in the other lambda clones and therefore one of these two clones, CL#034, was selected for further analysis.

In summary, the seven positive lambda clones analysed are not identical since they can be organised into at least four groups based on the DNA fragment patterns obtained by digestion with *Eco*RI and *Bam*HI. However, all seven lambda clones possess inserts that overlap one another, including a 4 kb *Eco*RI fragment that hybridises to the *cpr-3* cDNA clone (cm12b6). These data suggest that all seven overlapping lambda clones are derived from a single locus and possess inserts that contain the *cpr-3* gene. A single 4 kb hybridising fragment was also observed when Southern blots of *Eco*RI digested *C.elegans* Bristol N2 genomic DNA and yeast DNA containing the Y69A3 YAC clone were probed with the cm12b6 cDNA clone of *cpr-3* (Figure 3.1). These data suggest that all seven lambda clones are derived from the same locus as that covered by the Y69A3 YAC clone, supporting the conclusion that there is a single copy of the *cpr-3* gene within the *C.elegans* genome. One of these lambda clones, CL#034 was selected for further analysis.

3.2.8.4. Obtaining subclones containing *cpr-3* from CL#034

DNA from the lambda clone CL#034 was digested with a variety of restriction endonucleases in both single and double digests. The digests were electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N filters (Amersham International plc) and probed with the *cpr-3* cDNA clone (cm12b6). This analysis identified a 5 kb *Xho*I/*Sac*II fragment that hybridised to cm12b6 (Figure 3.15, lane 6) and this fragment was selected for subcloning. CL#034 DNA was digested with *Xho*I and *Sac*II and electrophoresed through a 1.0% agarose gel. The 5 kb *Xho*I/*Sac*II fragment could not be resolved from two slightly larger fragments. Thus, all three bands were gel purified and ligated into *Xho*I/*Sac*II digested, gel purified pBluescript II KS+. The ligation mix was transformed into *E.coli* XL1-blue and positive clones were selected using ampicillin resistance and blue/white colony selection. Plasmid DNA was prepared from overnight cultures of six selected clones. The plasmid DNA was digested with *Xho*I and *Sac*II, electrophoresed through a 0.7% agarose gel, Southern blotted to Hybond N (Amersham International plc) and probed with the *cpr-3* cDNA clone (cm12b6). Analysis of the ethidium bromide stained agarose gel prior to Southern blotting indicated that all six subclones contained single inserts of the correct size. The autoradiographic results indicated that the insert of all six selected subclones hybridised to the *cpr-3* cDNA clone. Thus one of the subclones was selected for further analysis and named CL#034/5/KS+.

3.2.9. Restriction endonuclease mapping of the genomic clones

As a first step in restriction mapping, single digests were performed to determine suitable enzymes for mapping each subclone. The digests for each subclone were electrophoresed through 0.6% agarose gels and the bands visualised by ethidium bromide staining. Restriction endonucleases which cut the inserts at three or less sites were selected for restriction mapping.

For restriction mapping, a variety of single and double digests were performed on each of the four subclones. The digested DNA was electrophoresed through 0.7% agarose gels, the DNA fragment patterns generated were visualised by ethidium bromide staining and were recorded on Polaroid film. Southern blots were prepared for each agarose gel using Hybond N (Amersham International plc), probed with the intact cDNA appropriate for each subclone and washed at high stringency. After exposure to medical X-ray film; the probed Southern blots were stripped, reprobed with the 5′ end of each cDNA clone, washed at high stringency and exposed to medical X-ray film. The 5′ end cDNA probes were generated by PCR using the M13 Forward primer and a primer specific to each cDNA clone (Chapter 2, Section 2.13.3) The PCR reaction yielded products of the expected size, 180 bp for cm12b6, 145 bp for cm14e3, 190 bp for cm04d10 and 230 bp for cm01a5. These PCR products were gel purified and radiolabeled by random priming.

Figures 3.16, 3.17, 3.18 and 3.19 show the results of the digests for each of the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, obtained by ethidium bromide staining of the agarose gels prior to Southern blotting. Each figure also includes a table that summarises these results, in addition to summarising the results obtained from probing the Southern blots with the intact cDNA and 5′ end cDNA probes. The final restriction maps of the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, as well as the DNA fragments used to generate these maps, are shown in Figures 3.20, 3.21, 3.22 and 3.23. The position, orientation and maximum length of the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* within the appropriate subclone, determined from the Southern blot results using the appropriate intact cDNA and 5′ end cDNA probes, are also shown in these figures.

For all four subclones, the restriction maps were recreated using the smaller fragments generated by the digests because the sizes of these fragments could be determined more accurately. The total insert sizes obtained from the restriction map data for the four subclones are very similar, but not identical, to the insert sizes obtained by Southern blot analysis prior to generating the subclones. Thus, Southern blot analysis predicts the sizes of the DNA fragments used to generate the subclones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to be 5 kb (Figure 3.15), 6.8 kb (Figure 3.12), 11 kb (Figure 3.12) and 10.7 kb (Figure 3.12), while the restriction map data of these four subclones generates sizes of 4.85 kb, 7 kb, 10.65 kb and 10.5 kb, respectively. The sizes generated from the restriction map data are likely to more accurately reflect the true insert sizes of these four subclones than the Southern blot data, since multiple smaller fragments, whose sizes could be more accurately determined, were used to recreate these restriction maps.

For the two larger subclones, W02B2/11/KS- (Figure 3.22) and C25B8/10.7/KS- (Figure 3.23), some of the restriction sites could not be aligned as accurately as those determined for CL#034/5/KS+ and ZK1055/6.8/KS-, resulting in slight discrepancies in the positions of certain restriction sites. This is assumed to be a result of experimental error, caused by the limits of resolution of the agarose gels used to determine the DNA fragment sizes required for creating the restriction maps. Indeed, discrepancies mostly arose when restriction sites were predicted from very large fragments, whose sizes could not be determined as accurately. It is not possible to discount the presence of additional restriction sites for the restriction endonucleases used during the mapping experiments which are in close proximity to those sites already determined for these enzymes. Such sites might result in the generation of very small restriction fragments which would not be detectable on the ethidium bromide stained agarose gels. These fragments would necessarily be very small and might therefore also account for the small discrepancies in the insert sizes predicted from the restriction map data and the Southern blot data.

For each of the four subclones, the high degree of similarity in the total insert size, obtained from the restriction map data, to the observed insert size, obtained from Southern blot analysis of the cosmid or lambda clones prior to subcloning, indicates that these subclones do not contain large tandem repeats. Such tandem repeats would result in multiple fragments of the same size and cause a severe underestimation of the insert size predicted by the restriction maps.

3.2.9.1. The position, orientation, maximum size and copy number of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* within their subclones

The Southern blot data obtained using the intact cDNA clones as probes revealed the position of each of the four genes within the four subclones, while the Southern blot data obtained using the 5′ end cDNA probes revealed the orientation of each of the four genes within the appropriate subclone. These results are summarised in Figures 3.20, 3.21, 3.22 and 3.23. Unfortunately, these results indicate that the 5′ end of the *cpr-3* gene is absent in the CL#034/5/KS+ subclone (Figure 3.20). Thus it was necessary to obtain a second subclone containing the 5′ end of the *cpr-3* gene and this is described in Section 3.2.10. Later nucleotide sequencing of the *cpr-3* gene revealed a *Sac*II site 116 bp downstream of the start of the putative translation initiation codon. This site was used for subcloning the insert of the CL#034/5/KS+ clone and therefore explains the absence of the 5′ end of the *cpr-3* gene within this clone. The nucleotide sequence data also explain why the 5′ end of the *cpr-3* gene was not detected in the Southern blot experiments performed prior to subcloning the insert of CL#034/5/KS+ (Figure 3.24). The cDNA clone of *cpr-3* (cm12b6) and the 5′ end of the *cpr-3* gene share only 111 bp of common nucleotide sequence upstream of the *Sac*II site and this region is interrupted by an intron (Figure 3.24). Thus, there is probably insufficient homology between the 5′ ends of the *cpr-3* cDNA clone (cm12b6) and the *cpr-3* gene to produce a hybridising band with the high wash stringencies used.

The restriction map data for each of the four subclones also indicate the maximum size of each gene, assuming that the cDNA clones that were used as probes are nearly full length. This was later demonstrated by sequencing the cDNA clones and conducting 5′ Rapid Amplification of cDNA Ends (5′ RACE) experiments (Chapter 4, Sections 4.2.2 and 4.2.3). According to the restriction map data, the maximum gene sizes of, *cpr-4*, *cpr-5* and *cpr-6*, contained within the subclones ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-, are 2.6 kb, 3 kb and 3.7 kb respectively. The maximum length of the portion of the *cpr-3* gene contained within the CL#034/5/KS+ subclone is 1.7 kb. Later sequencing of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* revealed their sizes, from the ATG initiation codon to the translation termination codon, to be 1611 bp, 1058 bp, 1269 bp and 1719 bp respectively. For *cpr-3*, 1493 bp of the gene, from the *Sac*II site to the translation termination codon, were contained within the CL#034/5/KS+ subclone. These results indicate that the hybridising regions of ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- are potentially large enough to contain more than one copy of *cpr-4*, *cpr-5* and *cpr-6* respectively. However, this is not the case since the 5′ end cDNA probes of these three genes hybridised to DNA fragments from the appropriate subclone which are too small to contain more than one copy of the gene. Thus the 5′ end cDNA probes for *cpr-4*, *cpr-5* and *cpr-6* hybridised to DNA fragments of 600 bp, 1400 bp and 700 bp, respectively. Of these fragments, only the 1400 bp *Bam*HI/*Xba*I DNA fragment from W02B2/11/KS-, which hybridised to the 5′ end cDNA probe of *cpr-5*, is potentially large enough to contain a second copy of this gene. However, later sequencing of the *cpr-5* gene revealed that over 700 bp of this gene is contained within this fragment, leaving insufficient room for a second copy. Thus all four subclones contain only a single copy of the appropriate gene as expected from previous data which suggest that each of these genes are present as single copies within the genome of *C.elegans*.

3.2.9.2. The nucleotide sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* support the restriction mapping data

Sequencing of the *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (Chapter 4) revealed restriction endonuclease sites within these four genes that were in complete agreement with the results obtained from the restriction mapping experiments and these results are summarised in Figures 3.25, 3.26, 3.27 and 3.28. It should be noted that these figures contain additional information about subclones generated later. As discussed in Section 3.2.9.1, the nucleotide sequence of *cpr-3* explained why the 5′ end of this gene was truncated in the CL#034/5/KS+ clone and also explained why the 5′ end of this gene was not detected by the Southern blot experiments performed prior to subcloning the insert of CL#034/5/KS+.

The restriction map of CL#034/5/KS+ (summarised in Figure 3.25) predicts the presence of a *Sac*II site, two *Hin*dIII sites, a *Cla*I site and a *Pst*I site in that order within the *cpr-3* gene. Nucleotide sequencing of the *cpr-3* gene confirmed the presence of these restriction endonuclease sites in the same order as predicted from the restriction map. Furthermore, the nucleotide sequence of *cpr-3* reveals that these sites are separated from one another by approximately the same number of nucleotides as expected from the restriction map. Thus, nucleotide sequencing indicates that the first *Hin*dIII site is 259 bp downstream of the *Sac*II site, the second *Hin*dIII site is 911 bp downstream of the first *Hin*dIII site, the *Cla*I site is 350 bp downstream of the second *Hin*dIII site and the *Pst*I site is 194 bp downstream of the *Cla*I site, and the restriction map data predicted these sites to be separated by 250 bp, 900 bp, 400 bp and 150 bp, respectively. In addition, the restriction map data suggest that no *Xba*I, *Eco*RI or *Xho*I restriction endonuclease sites are present in the *cpr-3* gene and this is also confirmed by the nucleotide sequence of *cpr-3*.

The restriction map of ZK1055/6.8/KS- (summarised in Figure 3.26) predicted the presence of a single *Bam*HI site within the *cpr-4* gene and a single *Kpn*I site approximately 600 bp upstream of the *Bam*HI site, possibly within the 5′ flank of *cpr-4*. The nucleotide sequence of *cpr-4* reveals that this is the case, by identifying a single *Bam*HI site within the *cpr-4* gene (370 bp downstream of the start of the ATG translation initiation codon) and a single *Kpn*I site 516 bp upstream of the *Bam*HI site (in the 5′ flank of this gene). In addition, the restriction map data suggest that no *Pst*I, *Xba*I, *Hin*dIII or *Cla*I restriction endonuclease sites are present in the *cpr-4* gene and this is also confirmed by the nucleotide sequence of *cpr-4*.

The restriction map of W02B2/11/KS- (summarised in Figure 3.27) predicted the presence of a single *Bam*HI site within the *cpr-5* gene and this is confirmed by the nucleotide sequence of *cpr-5*, which identifies a single *Bam*HI site 711 bp downstream of the start of the ATG translation initiation codon. In addition, the restriction map data suggest that no *Xba*I, *Xho*I, *Sac*I, *Pst*I or *Cla*I restriction endonuclease sites are present in the *cpr-5* gene and this is also confirmed by the nucleotide sequence of *cpr-5*.

The restriction map of C25B8/10.7/KS- (summarised in Figure 3.28) predicted the presence of a single *Bam*HI site within the *cpr-6* gene and a single *Sal*I site approximately 700 bp upstream of the *Bam*HI site, possibly within the 5′ flank of this gene. The nucleotide sequence of *cpr-6* revealed that this is the case, by identifying a single *Bam*HI site within the *cpr-6* gene (725 bp downstream of the start of the ATG translation initiation codon) and a single *Sal*I site 734 bp upstream of the *Bam*HI site (9 bp upstream of the start of the ATG translation initiation codon). In addition, the restriction map data suggest that no *Sac*I, *Pst*I or *Kpn*I restriction endonuclease sites are present in the *cpr-6* gene and this is also confirmed by the nucleotide sequence of *cpr-6*.

In summary, the restriction maps generated for the four subclones, CL#034/5/KS+, ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- agree with the nucleotide sequence obtained for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* contained within these subclones. These data provide solid evidence for; the efficacy of each of the restriction maps, the single copy nature of each gene within the appropriate subclone and the correct positioning and orientation of each gene within the appropriate subclone.

3.2.10. Subcloning the 5′ flank of *cpr-3* from CL#034

The restriction map data of CL#034/5/KS+ (Figure 3.20) was used to select enzymes which might be useful for subcloning the 5′ end of the *cpr-3* gene. Mapped sites for *Hin*dIII, *Cla*I, *Xba*I, *Pst*I and *Eco*RI sites were available in the subclone CL#034/5/KS+. Therefore, the lambda clone CL#034 was digested with each of these enzymes, the products electrophoresed through a 0.8% agarose gel and Southern blotted to Hybond N (Amersham International plc). The blot was then probed with the 250 bp *Hin*dIII/*Sac*II fragment from CL#034/5/KS+. This fragment contained part of the 5′ end of *cpr-3* present within CL#034/5/KS+. The results of this (Figure 3.29) revealed that *Hin*dIII, *Cla*I and *Eco*RI generated fragments upstream of, but overlapping, the CL#034/5/KS+ clone. The 1.7 kb *Hin*dIII and 3 kb *Cla*I fragments were gel purified, subcloned into pBluescript II KS+ and transformed into *E.coli* XL1-blue. Plasmid DNA was extracted from overnight cultures of selected transformants. The cloned 1.7 kb *Hin*dIII fragment was digested with *Hin*dIII while the cloned 3 kb *Cla*I fragment was digested with both *Cla*I and *Hin*dIII. These data indicated that both fragments had been subcloned and therefore a clone from each was selected and named CL#034/1.7/KS+ and CL#034/3/KS+ respectively. The restriction map data of CL#034/5/KS+ suggested that both subclones CL#034/1.7/KS+ and CL#034/3/KS+ would contain approximately the same amount of 5′ flank sequence from the *cpr-3* gene. The *Cla*I/*Hin*dIII double digest of CL#034/3/KS+ indicated that it only contained approximately 75 bp more 5′ flank sequence than the 1.7 kb *Hin*dIII fragment of CL#034/1.7/KS+, and this was confirmed by sequencing into the 5′ end of CL#034/3/KS+ using the T3 primer. Though CL#034/3/KS+ provided a small amount of extra 5′ flank sequence useful for later *lacZ* fusion constructs, the 1.7 kb *Hin*dIII fragment could be more readily subcloned into the *lacZ* fusion vectors to be used, which contained a *Hind*III site in their polylinkers. Thus CL#034/1.7/KS+ was selected for later use in the generation *lacZ* fusion constructs for *cpr-3*.

3.2.11. The generation of subclones from ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS-

Subclones containing the 5′ flanks of the three genes *cpr-4*, *cpr-5* and *cpr-6* were generated from the subclones ZK1055/6.8/KS-, W02B2/11/KS- and C25B8/10.7/KS- respectively, for later use in the generation of *lacZ* fusion constructs. The restriction map data for ZK1055/6.5/KS-, W02B2/11/KS- and C25B8/10.7/KS- (Figures 3.21, 3.22 and 3.23) were used to determine the appropriate fragments for subcloning into pBluescript II KS+. The 2.8 kb *Bam*HI/*Hin*dIII fragment of ZK1055/6.8/KS-, the 3 kb *Bam*HI fragment of W02B2/11/KS-, the 3.7 kb *Bam*HI/*Pst*I fragment of

C25B8/10.7/KS- and the 3 kb *Bam*HI/*Sal*I fragment of C25B8/10.7/KS- were gel purified and ligated into pBluescript II KS+. The ligation products were transformed into *E.coli* XL1-blue and several transformants were selected for each of the subclones prepared. Plasmid DNA was prepared from overnight cultures of these transformants, digested with the appropriate restriction endonucleases and electrophoresed through 0.7% agarose gels to check the sizes of the inserts. One of each of the subclones containing the 2.8 kb *Bam*HI/*Hin*dIII fragment of ZK1055/6.8/KS-, the 3.7 kb *Bam*HI/*Pst*I fragment of C25B8/10.7/KS- and the 3 kb *Bam*HI/*Sal*I fragment of C25B8/10.7/KS- were selected for further analysis and named ZK1055/2.8/KS+, C25B8/3.7/KS+ and C25B8/3/KS+. Since the 3 kb *Bam*HI insert could ligate into the pBluescript II KS+ vector in one of two orientations, several clones containing inserts of the correct size were selected for further analysis. The orientation of the 3 kb *Bam*HI insert of W02B2/3/KS+ in these clones was determined by sequencing into the ends using the T7 and KS primers. Two clones were selected, each containing the insert in the opposite orientation, and were named W02B2/3/KS+#1 and W02B2/3/KS+#2.

The subclone W02B2/11/KS+ proved to be unstable in *E.coli* XL1-blue cells. The subclone was therefore transformed into *E.coli* SURE competent cells (Stratagene Ltd.) which are deficient in most of the *E.coli* recombination pathways. The subclone W02B2/11/KS+ proved to be slightly more stable in *E.coli* SURE cells, however deleted clones were still frequently isolated after extracting DNA from overnight cultures of these cells. In contrast the clone W02B2/3/KS+ showed no such instability in *E.coli* XL1-blue cells. Therefore, I decided to subclone a smaller fragment from W02B2/11/KS+, containing the 3′ end of the *cpr-5* gene. The restriction map of W02B2/11/KS+ (Figure 3.22) was used to determine the appropriate fragment to clone. The 1.6 kb *Pst*I/*Bam*HI fragment was chosen, gel purified and ligated into the pBluescript II KS+ vector. The ligation products were transformed into *E.coli* XL1-blue cells and several transformants were selected. Plasmid DNA was prepared from overnight cultures of these transformants, digested with the appropriate restriction endonucleases and electrophoresed through 0.7% agarose gels to check the sizes of the inserts. One clone which contained the correct size of insert was selected for further analysis and named W02B2/1.6/KS+. The cloning of the 1.6 kb *Pst*I/*Bam*HI fragment from W02B2/11/KS+ was not ideal since this fragment was adjacent to but did not overlap W02B2/3/KS+. Therefore, it was possible that some sequences might be lost if two *Bam*HI sites were present in close proximity. However, later sequence comparison of the cDNA clone cm04d10 with the *cpr-5* gene, which spans both subclones, revealed this is not the case. The relationships of the subclones generated for each of the four genes, including the restriction enzymes sites used for cloning and the positions of the appropriate genes (determined by restriction mapping and sequencing), are summarised in Figures 3.25, 3.26, 3.27 and 3.28.

3.2.12. Determining whether the subclones are contiguous with the *C.elegans* genome

It was important to assess whether or not the subclones generated were contiguous with the *C.elegans* genome, particularly for those subclones required for generating *lacZ* fusion constructs to study the tissue specific expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (see Chapter 5). Two screens were performed to assess whether the subclones containing *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were contiguous with the *C.elegans* genome. In the first screen, the subcloned inserts were analysed to determine whether they contained fragments from the cosmid and lambda vectors used to generate the original clones. In the second screen, diagnostic digests were performed with *C.elegans* Bristol N2 genomic DNA using restriction endonucleases whose recognition sites had been mapped within the various subclones. This analysis was performed to determine whether the DNA fragment sizes and fragment patterns that were expected from the subclones were generated by the digested *C.elegans* genomic DNA after hybridisation with the appropriate cDNA clone.

For the first screen, the subclones containing *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were digested with the appropriate restriction endonucleases to release the subcloned inserts. The digests were electrophoresed through 1.0% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with sequences homologous to the original cosmid and lambda vectors. The results of these experiments demonstrated that none of the inserts from the subclones containing *cpr-3*, *cpr-4* or *cpr-5* hybridised with the vector probes, indicating that the inserts from these subclones do not possess any of the original cosmid or lambda vector sequences. However, the insert of subclone

C25B8/10.7/KS-, containing the *cpr-6* gene, did produce a weak hybridising signal when probed with pBluescript sequences, in addition to the strong hybridising signal expected from the approximately 3 kb pBluescript II KS- vector of this subclone (Figure 3.30). The cosmid vector pJB8 was used to generate the cosmid clone C25B8 from which C25B8/10.7/KS- was generated. Since the pJB8 and pBluescript vectors are both derived from the pBR322 plasmid, the hybridisation of pBluescript sequences to the cloned 10.7 kb insert of C25B8/10.7/KS- suggests that some original pJB8 cosmid vector sequence is present in this insert. The cosmid pJB8 possesses a *Pst*I site approximately 700 bp from the *Bam*HI site which was originally used for cloning the inserts obtained by *Sau*3A partial digestion of *C.elegans* genomic DNA during preparation of the cosmid library. The insert of C25B8/10.7/KS- was cloned using *Pst*I and *Sac*I sites. Therefore, the results suggest that there are 700 bp of pJB8 sequence present in this subclone and in the subclones C25B8/3.7/KS+ and C25B8/3/KS+, which also contain inserts cloned via the same *Pst*I site (Figure 3.28). To confirm this, a single sequencing reaction was performed using the T3 primer and C25B8/3/KS+. The nucleotide sequence generated shares complete base identity to sequences adjacent to the *Pst*I site of pJB8, within the 700 bp region of this vector defined by the *Pst*I and *Bam*HI sites. These results indicate that C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS-, whose inserts were all cloned via the same *Pst*I site, possess approximately 700 bp of the pJB8 cosmid vector within their inserts.

Since the pJB8 cosmid vector sequence is upstream of the mapped position of the *cpr-6* gene, it was necessary to determine how much 5′ flank sequence of the *cpr-6* gene was contained within the three subclones C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS- downstream of the pJB8 vector sequence. In order to do this, new restriction endonuclease sites within the 5′ flank of *cpr-6*, between the gene and the pJB8 vector sequences, had to be determined to provide sites for diagnostic digests of *C.elegans* genomic DNA. Later sequencing of *cpr-6*, revealed several *Hin*dIII sites present within the gene. These data were used with the restriction map data to generate a partial restriction map for C25B8/3.7/KS+ (Figure 3.31A). *Hin*dIII single and *Hin*dIII/*Sal*I double digests were used to determine the position of the *Hin*dIII site upstream of the *Sal*I site (Figure 3.31B). These results indicate that there is a *Hin*dIII site 1.35 kb upstream of the *Sal*I site in this subclone as well as in C25B8/3/KS+ and

C25B8/10.7/KS-, which cover the same region of the *C.elegans* genome. This site was later used for diagnostic digests of *C.elegans* genomic DNA.

In the second screen, *C.elegans* genomic DNA was digested with restriction endonucleases, whose sites had been mapped in the subclones containing the four genes. The DNA was electrophoresed through 0.7% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the cDNA clone appropriate for the set of digests. In some cases, the sites recognised by the enzymes used had not been mapped to all the subclones for a given gene. In these instances, the subclones were also digested with the same restriction endonucleases, Southern blotted to Hybond N (Amersham International plc) and probed with the appropriate cDNA clone to allow comparison with the pattern of hybridising fragments produced by the digested *C.elegans* genomic DNA. The restriction endonucleases were chosen after consulting the restriction map data for each of the subclones summarised in Figures 3.25, 3.26, 3.27 and 3.28.

Figure 3.32 shows the results obtained for the diagnostic digests used to determine whether the CL#034/5/KS+ and CL#034/1.7/KS+ subclones, containing the *cpr-3* gene are contiguous with the *C.elegans* genome. The *Eco*RV restriction endonuclease was used in some of the diagnostic digests. Though an *Eco*RV site had been mapped in CL#034/1.7/KS+, it was unknown whether additional *Eco*RV sites were present in CL#034/5/KS+. Therefore, a triple digest was performed with CL#034/5/KS+ DNA using the *Eco*RV, *Sac*II and *Hin*dIII enzymes. The result obtained after agarose gel electrophoresis of the digest, Southern blotting and probing with the *cpr-3* cDNA clone (cm12b6) is shown in Figure 3.32A. The hybridisation pattern observed indicates that no *Eco*RV sites are present in this subclone, since the hybridisation pattern observed is consistent with cleavage of the insert DNA at the mapped *Hin*dIII and *Sac*II restriction sites and release of the remaining insert DNA from the vector by cleavage at the *Eco*RV site within the pBluescript II KS+ polylinker. The expected hybridisation patterns, resulting from triple digests of *C.elegans* genomic DNA using *Eco*RV as one of the enzymes, are based on this observation (Figure 3.32C). The hybridisation patterns generated by the diagnostic digests of *C.elegans* genomic DNA are shown in Figure 3.32B while the expected hybridisation patterns are shown Figure 3.32C, based on the restriction map data obtained for CL#034/5/KS+ and CL#034/1.7/KS+ (summarised in

Figure 3.25). The hybridisation patterns obtained from the diagnostic digests were entirely as expected and therefore indicate that the CL#034/5/KS+ and CL#034/1.7/KS+ subclones are contiguous with the *C.elegans* genome.

Figure 3.33A shows the results obtained for the diagnostic digests used to determine whether the ZK1055/6.8/KS- subclone containing the *cpr-4* gene is contiguous with the *C.elegans* genome. The expected hybridisation patterns for the various digests are also summarised (Figure 3.33B), based on the restriction map data for ZK1055/6.8/KS- (summarised in Figure 3.26). The hybridisation patterns obtained from the diagnostic digests are entirely as expected from the restriction map data and therefore indicate that the ZK1055/6.8/KS- subclone is entirely contiguous with the *C.elegans* genome.

Figure 3.34A shows the hybridisation patterns obtained for the diagnostic digests used to determine whether the W02B2/3/KS+ and W02B2/1.6/KS+ subclones, containing the *cpr-5* gene, are contiguous with the *C.elegans* genome. The larger W02B2/11/KS- subclone was not analysed because this clone was unstable and therefore all subsequent experiments were performed using the two smaller subclones. Figure 3.34B shows the expected hybridisation patterns for the digests shown in Figure 3.34A. These expected patterns were obtained from the restriction map data for W02B2/3/KS+ and W02B2/1.6/KS+ (summarised in Figure 3.27). Again, the hybridisation patterns observed from the diagnostic digests are as expected from the restriction map data for these two subclones and therefore indicate that W02B2/3/KS+ and W02B2/1.6/KS+ are contiguous with the *C.elegans* genome.

For the subclones containing the *cpr-6* gene (C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3.0/KS+), a slightly different approach was used. These clones are not entirely contiguous with the *C.elegans* genome as a result of the presence of pJB8 cosmid vector sequence within the inserts (discussed above). Therefore, the *Hin*dIII restriction endonuclease was used in diagnostic digests of *C.elegans* genomic DNA because an internal *Hin*dIII site, downstream of the pJB8 sequence had been mapped in the C25B8/3.7/KS- subclone (discussed above). However, it was not known whether additional *Hin*dIII sites might be present in the region of the insert of C25B8/10.7/KS- not covered by the smaller C25B8/3.7/KS- subclone and outwith the region sequenced (Figure 3.28). Thus, both the largest subclone (C25B8/10.7/KS-) and *C.elegans* genomic DNA were digested using single and double digests involving *Hin*dIII. The results obtained after agarose gel electrophoresis of the digests, Southern blotting and probing with the *cpr-6* cDNA clone (cm01a5) are shown in Figure 3.35A and Figure 3.35B, along with the expected fragment patterns (Figure 3.35C), based on the known *Hin*dIII sites from the restriction map and sequence data for C25B8/10.7/KS-. The hybridisation patterns from the digested *C.elegans* genomic DNA and the digested C25B8/10.7/KS- DNA are identical. Furthermore, the fragment sizes and fragment patterns are as expected from the *Hin*dIII sites determined by restriction mapping or sequencing (summarised in Figure 3.28) with the exception of the absence of a 0.15 kb hybridising fragment for *Hin*dIII/*Sal*I double digests of both *C.elegans* genomic DNA and C25B8/10.7/KS- DNA, which was due to this small fragment running off the end of the gel. However, an additional hybridising band of approximately 1.3 kb was also observed for *C.elegans* genomic DNA and C25B8/10.7/KS- DNA with *Hin*dIII single or *Hin*dIII/*Sal*I double digests. Analysis of the ethidium bromide stained agarose gel of the C25B8/10.7/KS- DNA digests prior to Southern blotting suggest the additional 1.3 kb hybridising band is a result of partial digestion. This conclusion is supported by previous Southern blots of C25B8 cosmid DNA digested with *Hin*dIII and hybridised with the *cpr-6* cDNA clone (cm01a5) which only generated the four expected bands (lane 8, Figure 3.11A). Furthermore, *Hin*dIII/*Bam*HI double digests of both C25B8/10.7/KS- DNA and *C.elegans* genomic DNA only produce the expected bands (lane 2 Figure 3.35A and 3.35B).

The results of these experiments indicate that a previously undetermined *Hin*dIII site is present 2 kb downstream of the most 3′ *Hin*dIII site identified from sequencing the *cpr-6* gene. This site is approximately 1.65 kb downstream of the putative translation termination codon of *cpr-6*. The identical hybridisation patterns observed between the digested subclone and the digested *C.elegans* genomic DNA indicate that the region of C25B8/10.7/KS- covered by these *Hin*dIII sites is contiguous with the *C.elegans* genome. Importantly, this indicates that the region of the 5′ flank of *cpr-6*, up to the mapped *Hin*dIII site 1.35 kb upstream of the *Sal*I, site is contiguous with the *C.elegans* genome and that this contiguity extends to approximately 1.65 kb downstream of the translation termination codon of *cpr-6*.

3.2.13. The physical map positions of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

As stated in the discussion to this chapter, there is good alignment between the physical and genetic maps of *C.elegans*. Thus, information regarding the position of genes obtained from the physical map of *C.elegans* can be related to the genetic map. The physical map positions of the three cosmid clones ZK1055, W02B2 and C25B8, containing *cpr-4*, *cpr-5* and *cpr-6* respectively, were already known prior to isolation of the genes from these cosmids. However, no physical map data were available for the CL#034 lambda clone and therefore this clone was sent to J.Sulston and A.Coulson at the Sanger Centre for physical mapping. According to their analysis, CL#034 (and therefore *cpr-3*) was found to map to the Y69A3 YAC clone as expected from the physical map data for the *cpr-3* cDNA (cm12b6). Another cathepsin B-like gene, *cpr-1*, previously isolated from *C.elegans* (Ray and McKerrow, 1992) also maps to this region. Restriction endonuclease fingerprinting performed by A.Coulson indicated that CL#034 almost entirely overlaps the lambda clone from which *cpr-1* was isolated, suggesting that these clones contain two distinct but very closely linked genes, or that *cpr-1* and *cpr-3* are the same gene. Later sequencing of the *cpr-3* gene confirmed that *cpr-1* and *cpr-3* are distinct genes, having different gene architectures and coding for proteins with only 57.4% amino acid identity (Chapter 4, Sections 4.2.5 and 4.2.7). Alignment of the physical and genetic maps reveal that *cpr-3* maps to the right end of the central region of chromosome V, between the *unc-76* gene (map position +7.02) and the stP18 restriction fragment length polymorphism (RFLP, map position +9.69).

Alignment of the physical and genetic maps reveals that *cpr-4* and *cpr-5* are both on chromosome V but are not closely linked to one another or to *cpr-1* and *cpr-3*. The *cpr-4* gene, contained in the ZK1055 cosmid clone, maps to the centre of chromosome V between the stP192 RFLP (map position +0.01) and the *snb-1* gene (map position +0.18). The *cpr-5* gene, contained in the W02B2 cosmid clone, maps to the left end of chromosome V between the two RFLPs, nP61 (map position -20.23) and nP62 (map position -19.54).

In contrast, alignment of the physical and genetic maps reveals that *cpr-6*, contained in the C25B8 cosmid clone, is not linked to *cpr*-1, *cpr-3*, *cpr-4* or *cpr-5*. This gene maps to the centre of the X chromosome between *sup-7* (map position -2.9) and *unc-6* (map position -2.18).

3.3. Discussion

The experiments described in this chapter had two purposes; to determine whether the genome of *C.elegans* encodes a multigene family of cathepsin B-like enzymes and to clone such genes for future characterisation and analysis. The four cathepsin B-like genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, were cloned using four previously isolated cDNA clones (Waterston *et al.,* 1992). Later sequencing of the four genes (reported in Chapter 4) demonstrated that these genes are cathepsin B-like. The three genes, *cpr-4*, *cpr-5* and *cpr-6*, were cloned using the physical map of *C.elegans* and the physical map data supplied with the cDNA clones. This approach enabled selected sets of cosmid clones to be screened on the basis of the positional information obtained from the physical map data. These screens resulted in the isolation of 3 cosmid clones, ZK1055, W02B2 and C25B8, which contained the 3 genes *cpr-4*, *cpr-5* and *cpr-6* respectively. The *cpr-3* gene was isolated by screening a *C.elegans* genomic DNA lambda EMBL4 library which yielded the clone CL#034, containing this gene. The four genes were subcloned and these subclones were subsequently restriction mapped.

The isolation of four cathepsin B-like genes from *C.elegans* indicates that this free living nematode species possesses a cathepsin B-like multigene family with at least five members, comprising the four genes reported here and the previously isolated *cpr-1* gene (Ray and McKerrow, 1992). The physical map data and results from Southern blot analyses of digested *C.elegans* DNA and yeast genomic DNA, containing the appropriate YAC clone, indicate that each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, are located at a distinct locus within the *C.elegans* genome. These Southern Blot analyses also indicate that the four genes are sufficiently diverged not to cross hybridise strongly to one another or to other sequences in the *C.elegans* genome. The restriction map data for the subclones containing these genes indicate that each of the genes is single copy within its respective subclone or subclones. Together, these data indicate that these four genes are single copy, are diverged from one another and are dispersed throughout the genome of *C.elegans* and therefore suggest that they have arisen from an ancient gene duplication event. The dispersed genome locations of these four genes may have enabled them to evolve independently since these genes would no longer be subjected to the homogenising influences normally associated with tandem arrays, such as gene conversion and unequal crossover events.

The approximate positions of the clones containing the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, within the genome were determined by alignment of the physical and genetic maps of *C.elegans*. Four of the genes are located on chromosome V but only two genes are closely linked, the *cpr-3* gene and the previously isolated *cpr-1* gene. Despite their close linkage, these genes do not appear to be tandem repeats since later sequencing indicates that these two genes are quite diverged, possessing different gene architectures and encoding proteins with only 57.4% identity. The other two genes present on chromosome V, *cpr-4* and *cpr-5*, are not closely linked to one another or to *cpr-1* and *cpr-3* and therefore clearly do not represent tandem repeats. The *cpr-6* gene, is not linked to the other four genes, being located on the X chromosome.

In addition to allowing the approximate location of the four genes to be determined, alignment of the physical and genetic maps also yields interesting candidate mutant phenotypes in the region of the genetic map covering the physical map location of these genes. For *cpr-4*, the *let-473*, *let-462* and *let-473* lethal mutations were identified which have early or early / mid larval lethal phenotypes. For *cpr-5*, the *let-326* lethal mutation which has an early larval lethal phenotype was identified. Both these genes have temporal expression patterns consistent with these mutant phenotypes since expression of both genes is induced during late embryogenesis and is elevated during the larval stages of development (Chapter 5). For *cpr-6*, the *flu-2* mutation was identified. This mutation results in reduced gut fluorescence and may be a result of defects in the catabolic pathways which occur in the gut. The *cpr-6* gene has a tissue specific expression pattern consistent with this mutation, since this gene appears to be expressed exclusively in the intestinal cells of *C.elegans* (Chapter 5). In order to determine whether these mutant phenotypes are indeed a result of mutations in the *cpr-4*, *cpr-5* or *cpr-6* genes, plasmid rescue experiments could be performed. With such experiments, *C.elegans* strains carrying the desired mutation are transformed with plasmids containing wild-type copies of the gene to be tested. The progeny are then screened to determine whether the plasmid is able to rescue the mutant phenotype.

Figure 3.1

Southern blot analysis of *C.elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y69A3, both digested with *Bam*HI or *Eco*RI

The digested *C.elegans* genomic DNA (5µg) and digested yeast genomic DNA (1µg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-3* cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

Lane 1: *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y69A3 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y69A3 digested with *Eco*RI

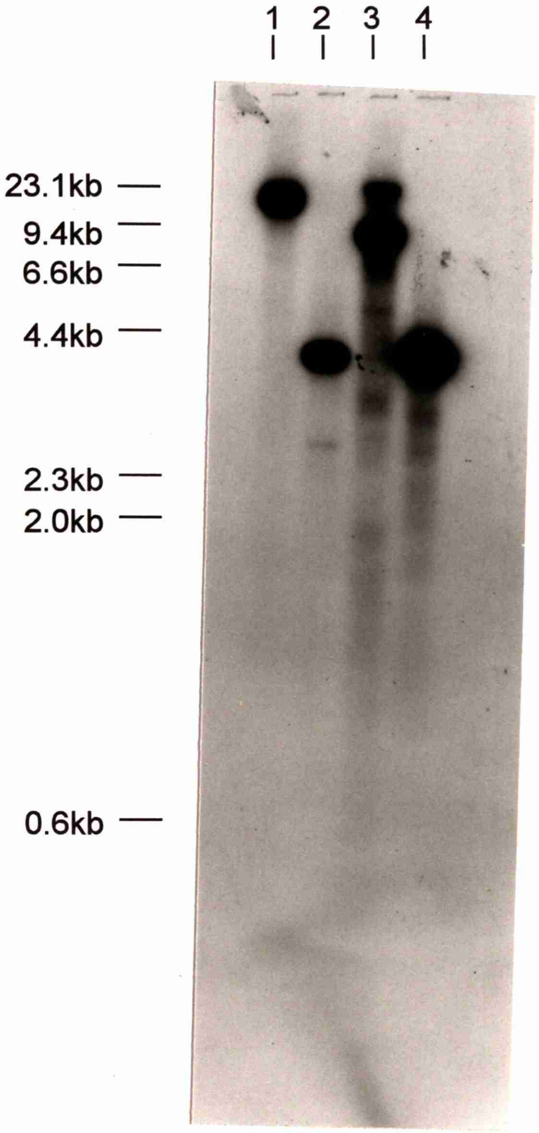

1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb

Figure 3.2

Southern blot analysis of *C.elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y7E11, both digested with *Bam*HI or *Eco*RI

The digested *C.elegans* genomic DNA (5µg) and digested yeast genomic DNA (1µg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-4* cDNA clone (cm14e3) and washed at high stringency prior to exposure to medical X-ray film.

Lane 1: *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y7E11 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y7E11 digested with *Eco*RI

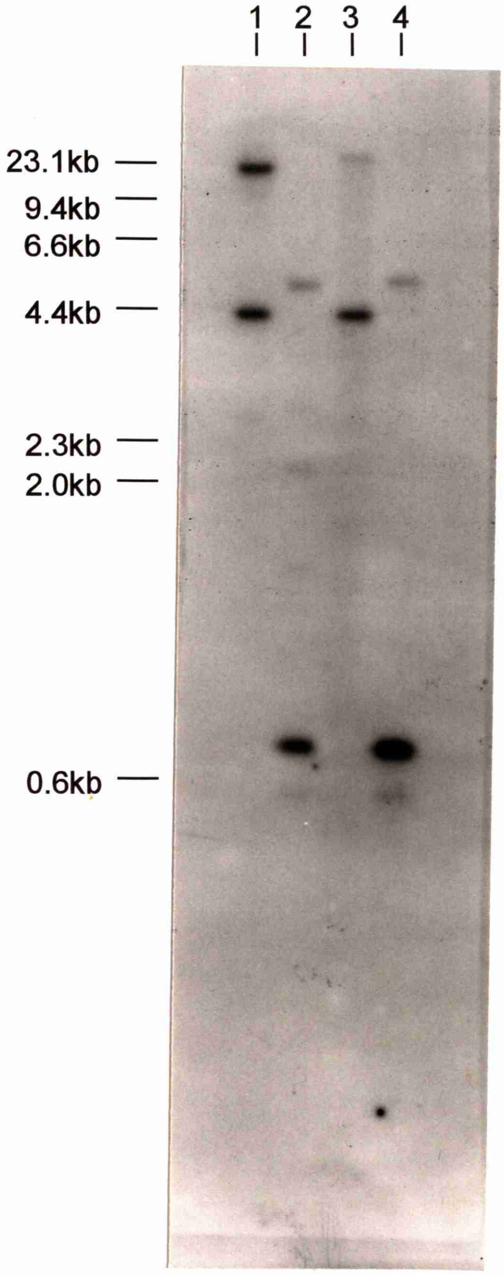

1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb

Figure 3.3

Southern blot analysis of *Bam*HI or *Eco*RI digested *C.elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y50D4 (**A**) or Y48G5 (**B**)

The digested *C.elegans* genomic DNA (5μg) and digested yeast genomic DNA (1μg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-5* cDNA clone (cm04d10) and washed at high stringency prior to exposure to medical X-ray film.

A. **Lane 1:** *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y50D4 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y50D4 digested with *Eco*RI

B. **Lane 1:** *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y48G5 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y48G5 digested with *Eco*RI

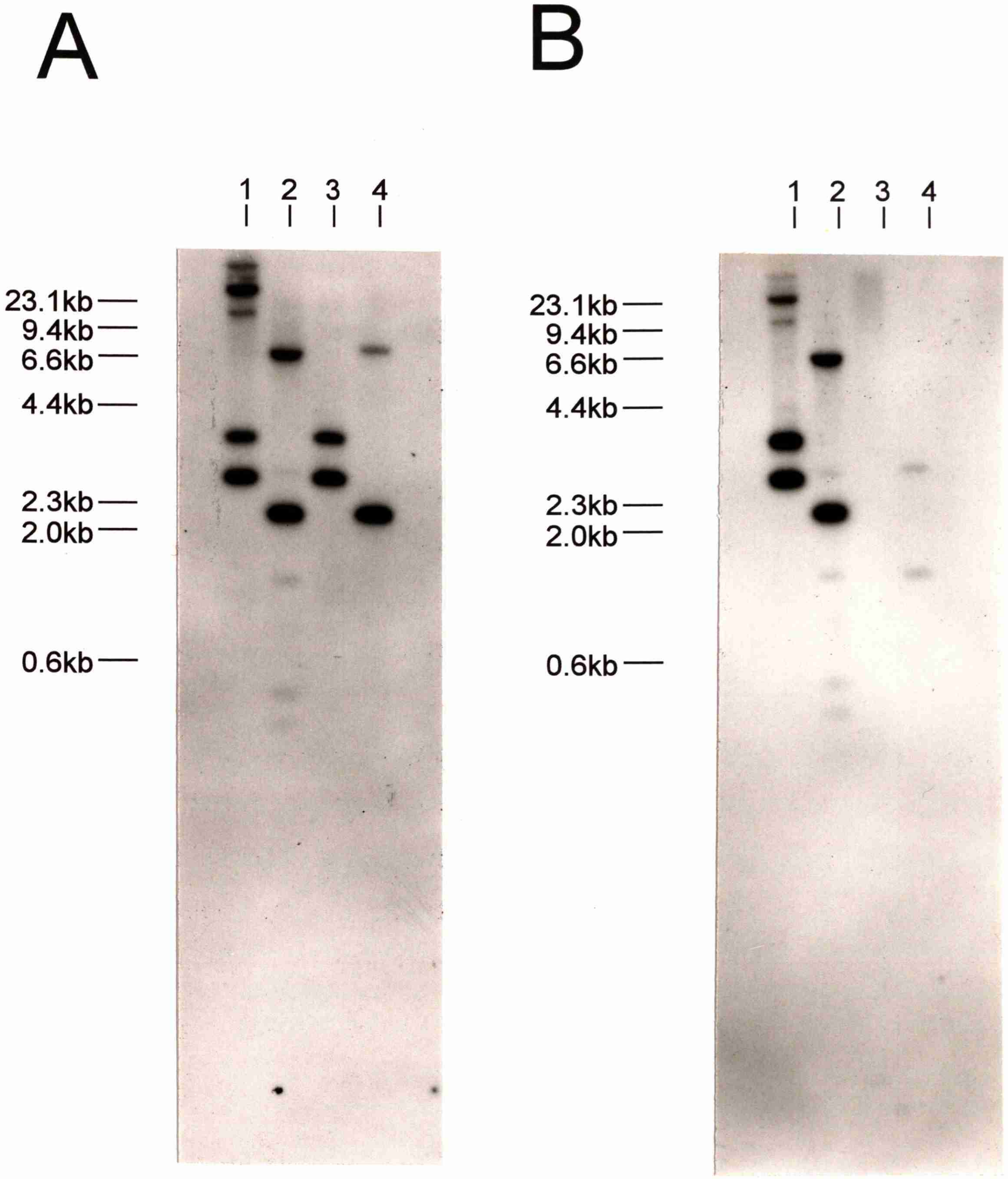

A
B
1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb
1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb

Figure 3.4

Southern blot analysis of *Bam*HI or *Eco*RI digested *C.elegans* genomic DNA and yeast genomic DNA, containing the YAC clone Y55B10 (A) or Y43B9 (B)

The digested *C.elegans* genomic DNA (5µg) and digested yeast genomic DNA (1µg) were electrophoresed through a 0.8% agarose gel, Southern blotted to Hybond N (Amersham International plc), hybridised with the *cpr-6* cDNA clone (cm01a5) and washed at high stringency prior to exposure to medical X-ray film.

A. **Lane 1:** *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y55B10 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y55B10 digested with *Eco*RI

B. **Lane 1:** *C.elegans* genomic DNA digested with *Bam*HI

Lane 2: *C.elegans* genomic DNA digested with *Eco*RI

Lane 3: yeast genomic DNA containing Y43B9 digested with *Bam*HI

Lane 4: yeast genomic DNA containing Y43B9 digested with *Eco*RI

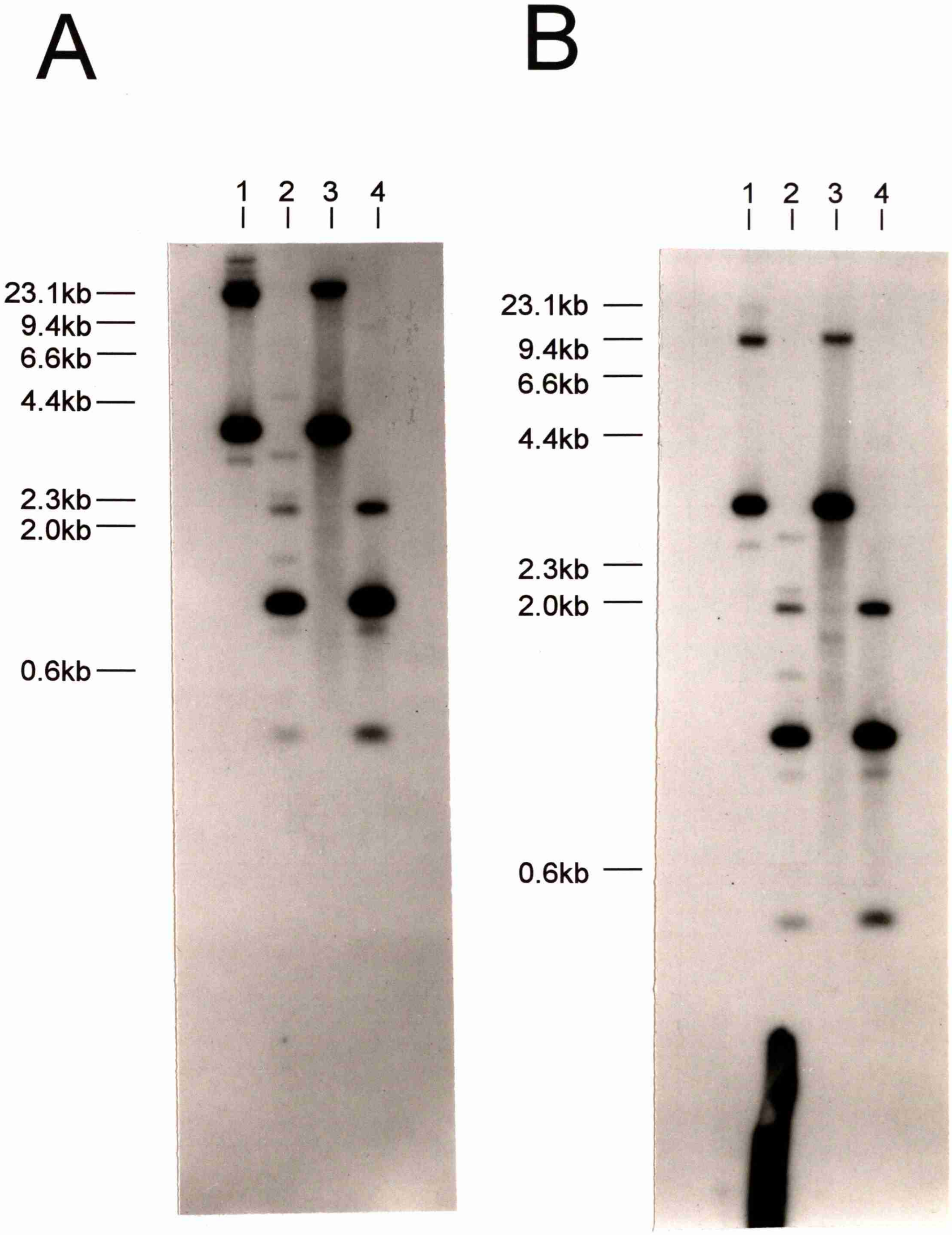

A
B
1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb
1
2
3
4
23.1kb
9.4kb
6.6kb
4.4kb
2.3kb
2.0kb
0.6kb

Figure 3.5

A summary of the physical map data for the *cpr-3* cDNA clone (cm12b6) supplied by R.Waterston

The figure shows the YAC clones covering the two regions of the YAC grid to which cm12b6 hybridised.

A. The region of the YAC grid covered by Y46B6.

B. The region of the YAC grid covered by Y69A3.

A and **B** do not show all the physical map data available for each region. They only show the data that were used to select cosmid clones which might contain the *cpr-3* gene and the cosmid clones that were selected. **A** and **B** were redrawn from information obtained from ACEDB.

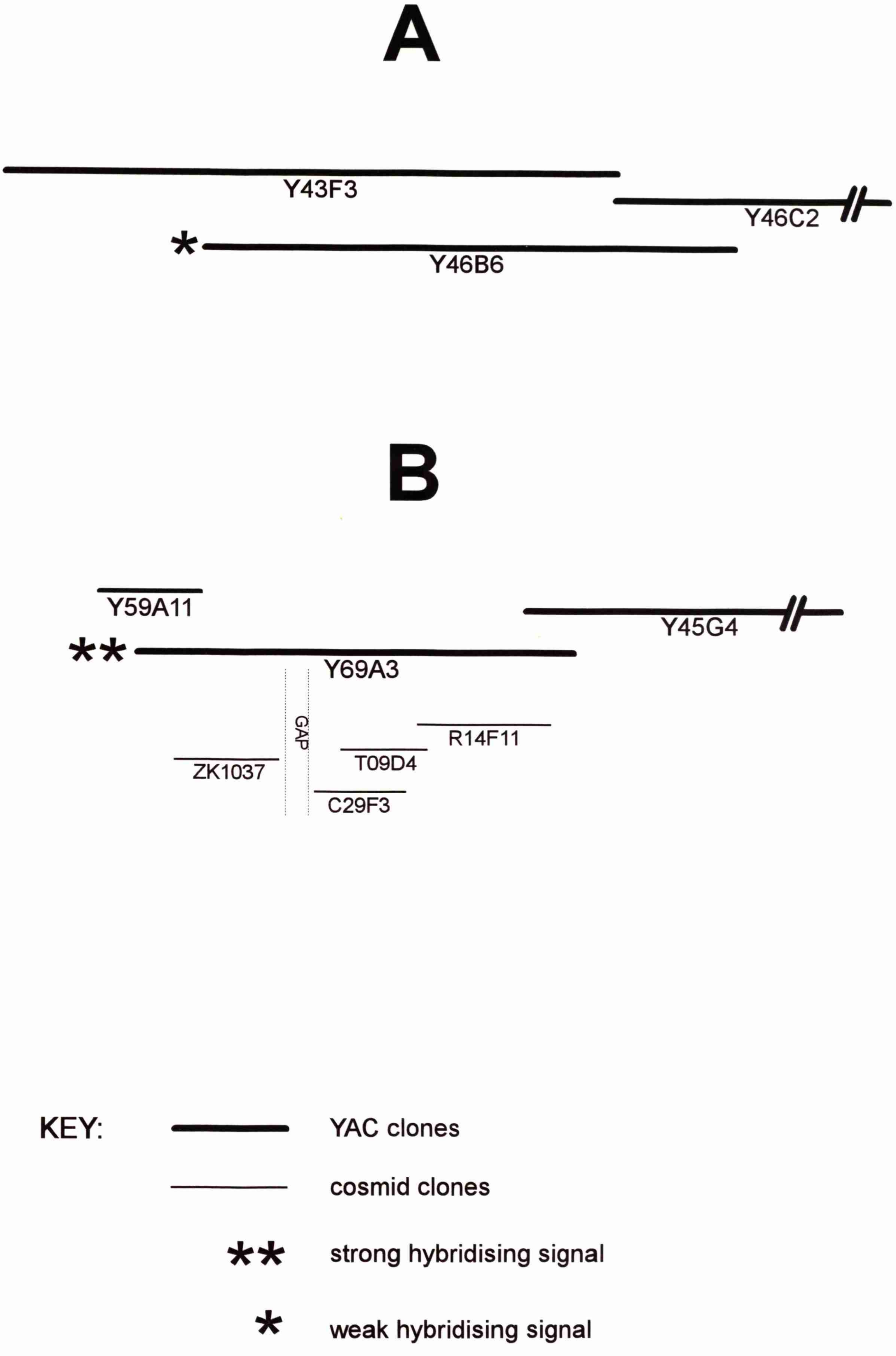

A
Y43F3
Y46C2
*
Y46B6
B
Y59A11
Y45G4
**
Y69A3
GAP
R14F11
T09D4
ZK1037
C29F3
KEY:
YAC clones
cosmid clones
** strong hybridising signal
* weak hybridising signal

Figure 3.6

A summary of the physical map data for the *cpr-4* cDNA clone (cm14e3) supplied by R.Waterston.

The figure shows the YAC clones covering the single region of the YAC grid to which cm14e3 hybridised. The figure does not show all the physical map data available for each region. It only shows the data that were used to select cosmid clones which might contain the *cpr-4* gene and the cosmid clones that were selected. The figure was redrawn from information obtained from ACEDB.

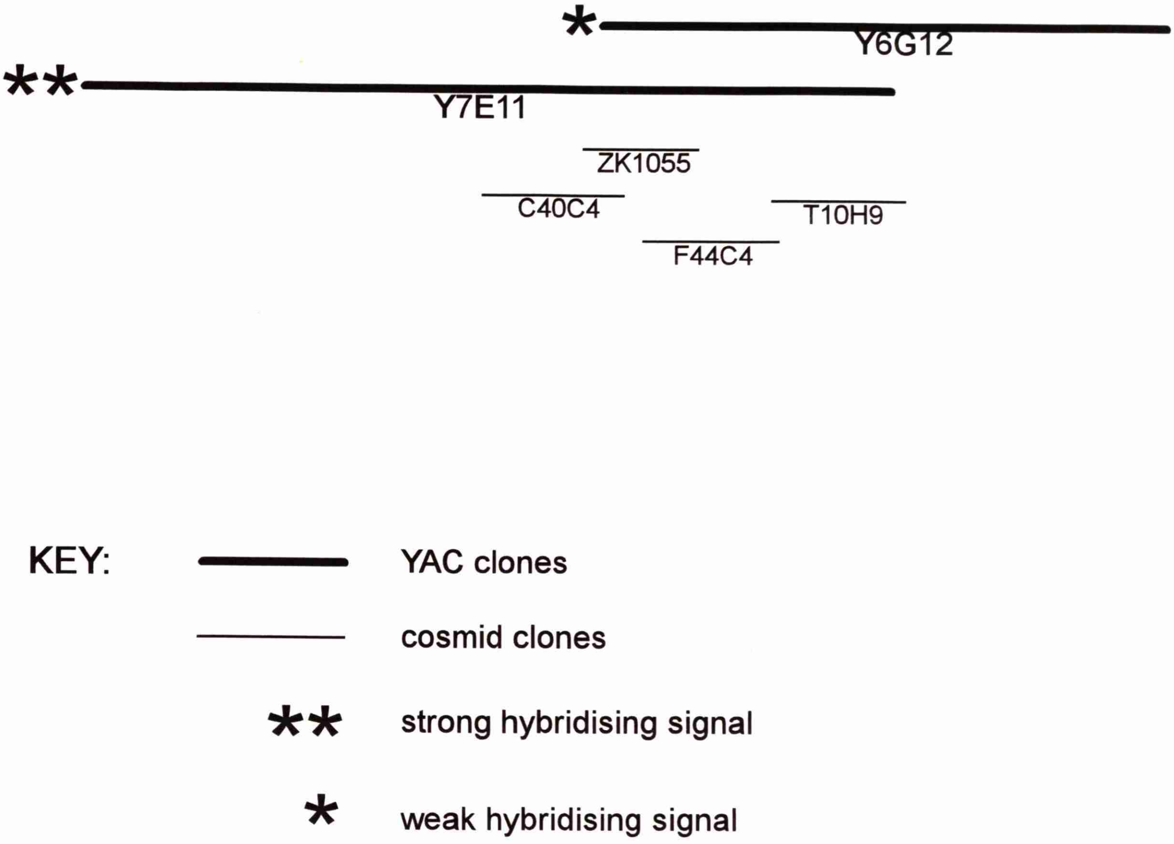

*
Y6G12
**
Y7E11
ZK1055
C40C4
T10H9
F44C4
KEY:
YAC clones
cosmid clones
**
strong hybridising signal
*
weak hybridising signal

Figure 3.7

A summary of the physical map data for the cm5d10 cDNA clone supplied by R.Waterston

This cDNA clone is assumed to be derived from the *cpr-5* gene because partial sequencing revealed that cm5d10 shares 98.7% nucleotide sequence identity with the cm04d10 cDNA clone used to isolate this gene. The figure shows the YAC clones covering the two regions of the YAC grid to which cm5d10 hybridised.

A. The region of the YAC grid covered by Y50D4.

B. The region of the YAC grid covered by Y48G5.

A and **B** do not show all the physical map data available for each region. They only show the data that were used to select cosmid clones which might contain the *cpr-3* gene and the cosmid clones that were selected. **A** and **B** were redrawn from information obtained from ACEDB.

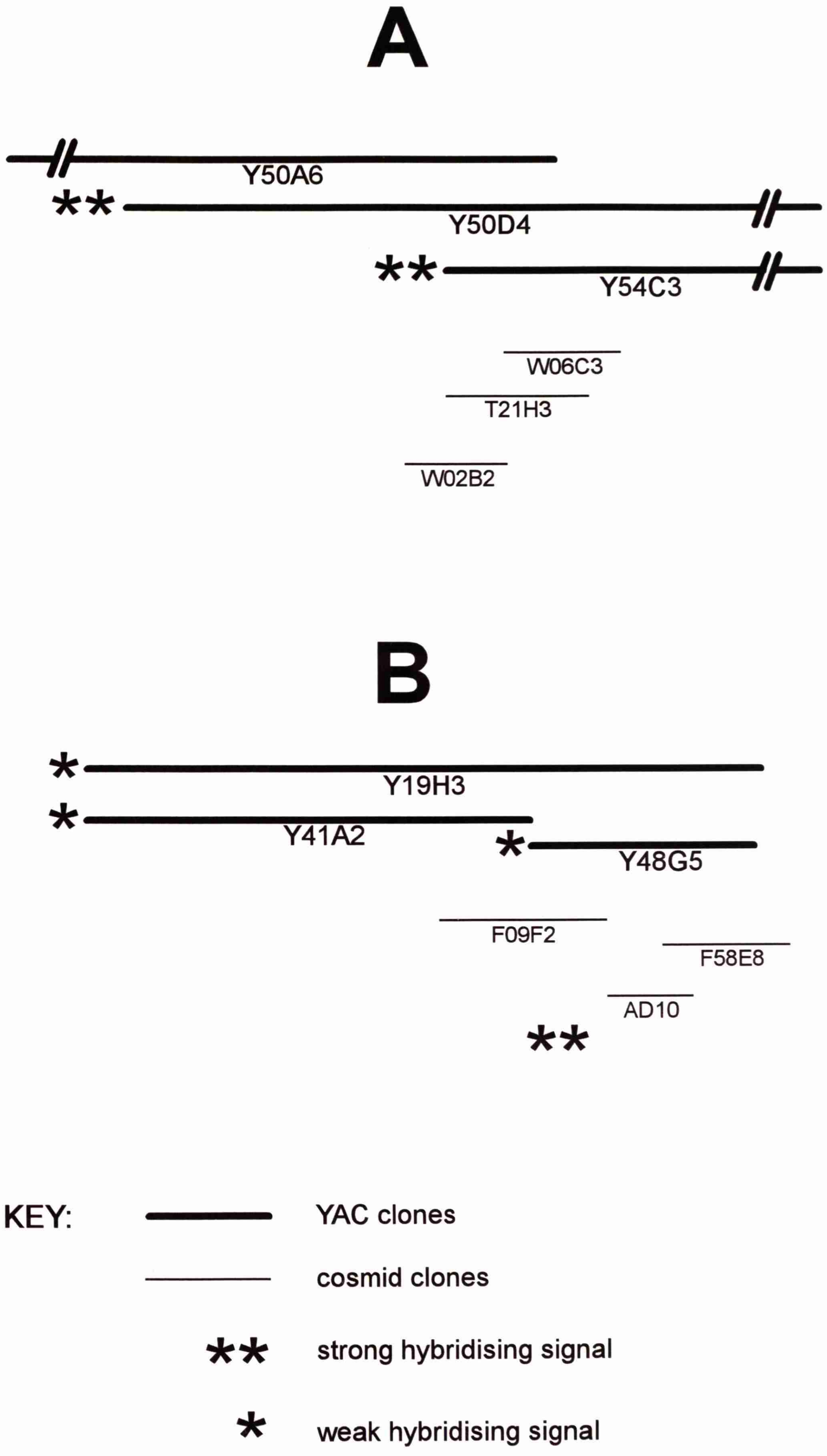

A
Y50A6
**
Y50D4
**
Y54C3
W06C3
T21H3
W02B2
B
*
Y19H3
*
Y41A2
*
Y48G5
F09F2
F58E8
AD10
**
KEY:
YAC clones
cosmid clones
** strong hybridising signal
* weak hybridising signal

Figure 3.8

A summary of the physical map data for the *cpr-6* cDNA clone (cm01a5) supplied by R.Waterston

The figure shows the YAC clones covering the single region of the YAC grid to which cm01a5 hybridised. The figure does not show all the physical map data available for each region. It only shows the data that were used to select cosmid clones which might contain the *cpr-6* gene and the cosmid clones that were selected. The figure was redrawn from information obtained from ACEDB.

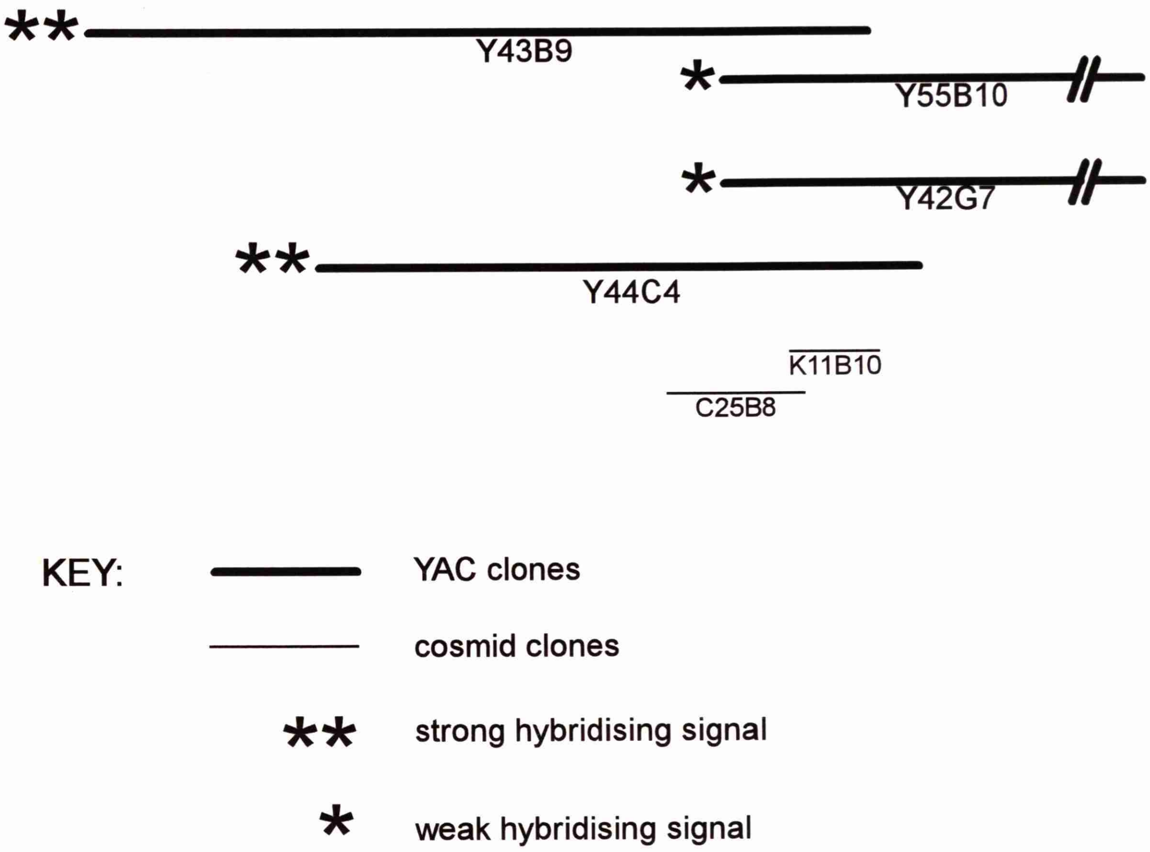

**
Y43B9
*
Y55B10
*
Y42G7
**
Y44C4
K11B10
C25B8
KEY:
YAC clones
cosmid clones
**
strong hybridising signal
*
weak hybridising signal

Figure 3.9

The results of Southern blot analysis of ZK1055 and F44C4 cosmid DNA

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the *cpr-4* cDNA clone (cm14e3). The filters were washed at high stringency prior to exposure with Medical X-ray film.

A. ZK1055 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

Lane 6	Lane 7	Lane 8	Lane 9	Lane 10	Lane 11
*Sac*II	*Xho*I	*Hind*III	*Spe*I	*Kpn*I	*Cla*I

B. F44C4 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

Lane 6	Lane 7	Lane 8	Lane 9	Lane 10	Lane 11
*Sac*II	*Xho*I	*Hind*III	*Spe*I	*Kpn*I	*Cla*I

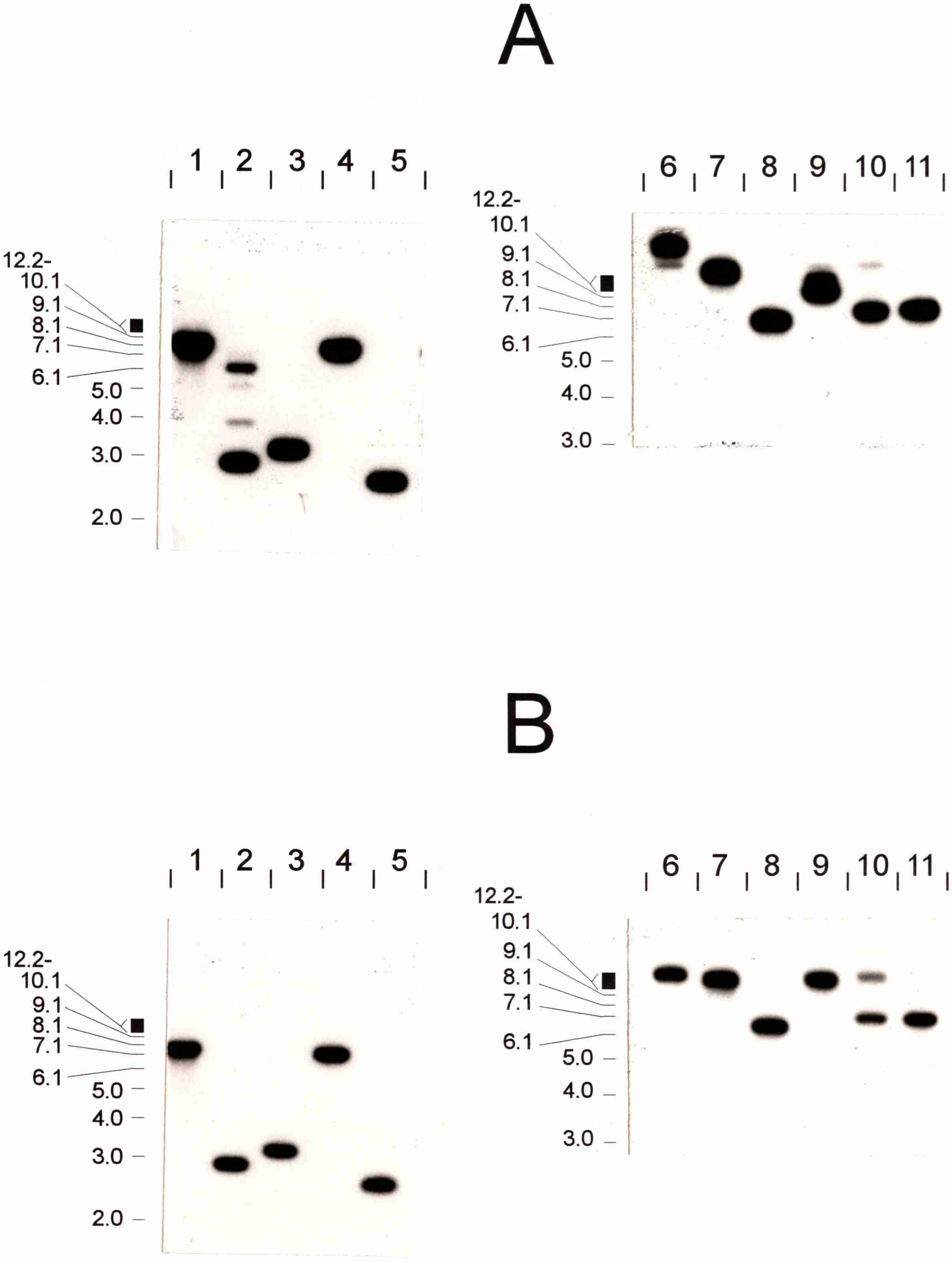

A
1
2
3
4
5
6
7
8
9
10
11
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
■
B
1
2
3
4
5
6
7
8
9
10
11
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
■

Figure 3.10

The results of Southern blot analysis of W02B2 and AD10 cosmid DNA

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the *cpr-5* cDNA clone (cm04d10). The filters were washed at high stringency prior to exposure with Medical X-ray film.

A. W02B2 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

Lane 6	Lane 7	Lane 8	Lane 9	Lane 10	Lane 11
*Sac*II	*Xho*I	*Hind*III	*Spe*I	*Kpn*I	*Cla*I

B. AD10 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

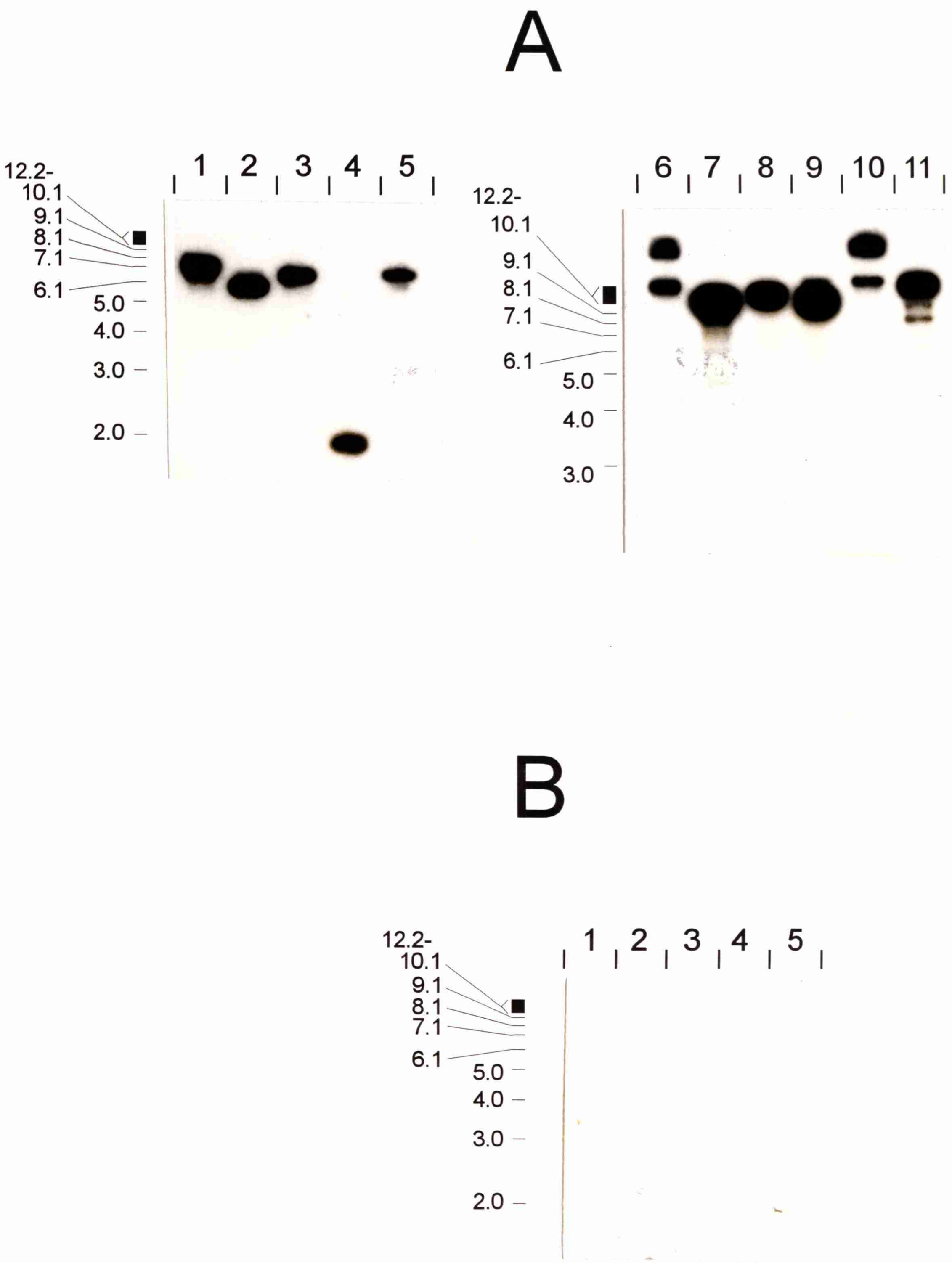

A
1 2 3 4 5
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
6 7 8 9 10 11
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
B
1 2 3 4 5
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0

Figure 3.11

The results of Southern blot analysis of C25B8 and K11B10 cosmid DNA

Cosmid DNA was digested with a variety of restriction endonucleases. The digests were electrophoresed through 0.8% agarose gels, Southern blotted to Hybond N (Amersham International plc) and hybridised with the *cpr-6* cDNA clone (cm01a5). The filters were washed at high stringency prior to exposure with Medical X-ray film.

A. C25B8 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

Lane 6	Lane 7	Lane 8	Lane 9	Lane 10	Lane 11
*Sac*II	*Xho*I	*Hind*III	*Spe*I	*Kpn*I	*Cla*I

B. K11B10 DNA digests

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5
*Apa*I	*Xba*I	*Sac*I	*Sal*I	*Pst*I

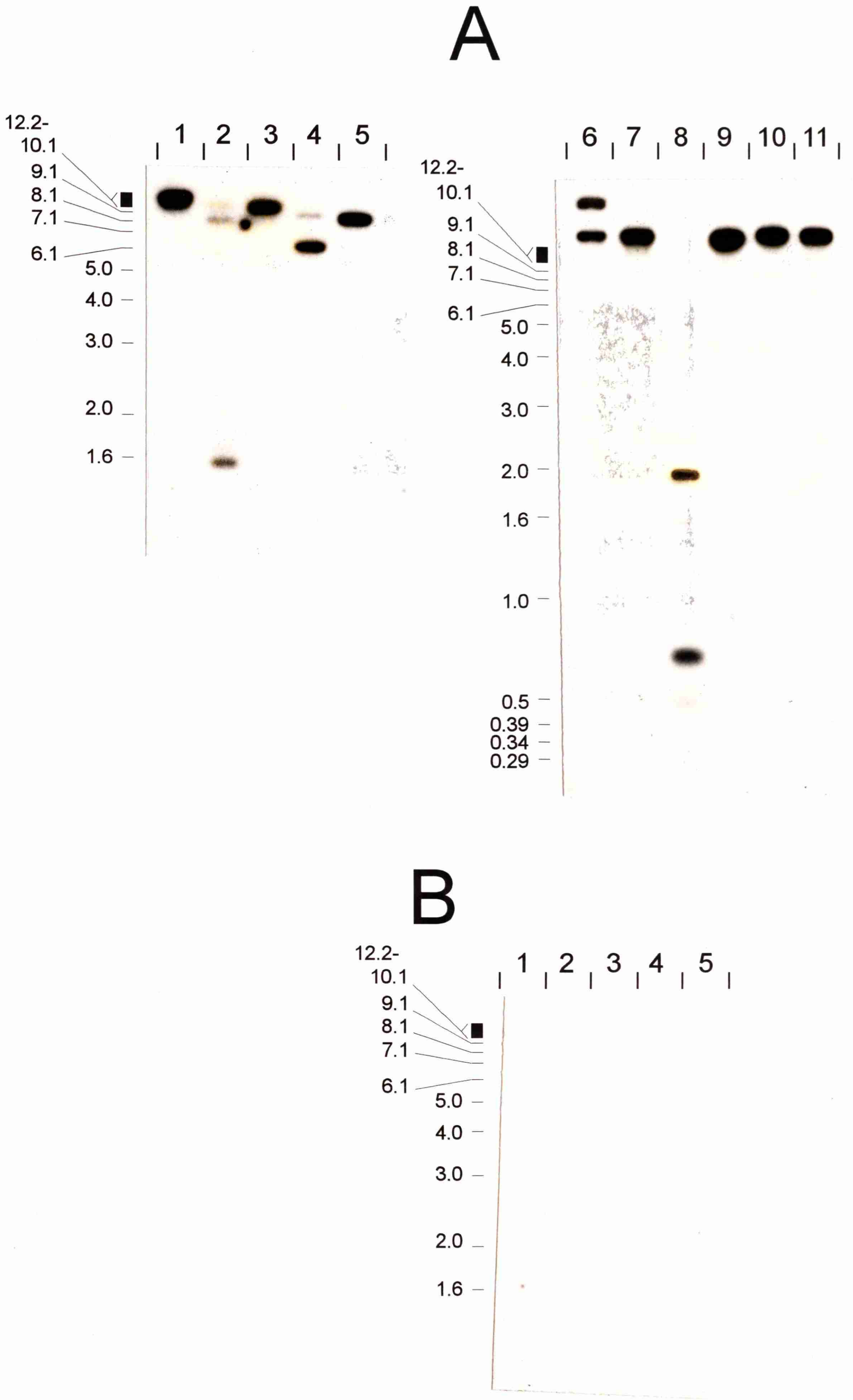

A
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
1.6
1
2
3
4
5
6
7
8
9
10
11
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
1.6
1.0
0.5
0.39
0.34
0.29
B
12.2-
10.1
9.1
8.1
7.1
6.1
5.0
4.0
3.0
2.0
1.6
1
2
3
4
5

Figure 3.12

Southern blots of restriction endonuclease digested ZK1055, W02B2 and C25B8 cosmid DNA

The restriction endonuclease digested cosmid DNA was electrophoresed through a 0.7% gel and Southern blotted to Hybond N (Amersham International plc). The blots were hybridised with the appropriate cDNA clone and washed at high stringency prior to exposure to medical X-ray film.

Lane 1: ZK1055 DNA digested with *Hin*dIII and probed with the *cpr-4* cDNA clone (cm14e3)

Lane 2: W02B2 DNA digested with *Pst*I and *Sac*I and probed with the *cpr-5* cDNA clone (cm04d10)

Lane 3: C25B8 DNA digested with *Pst*I and *Sac*I and probed with the *cpr-6* cDNA clone (cm01a5)

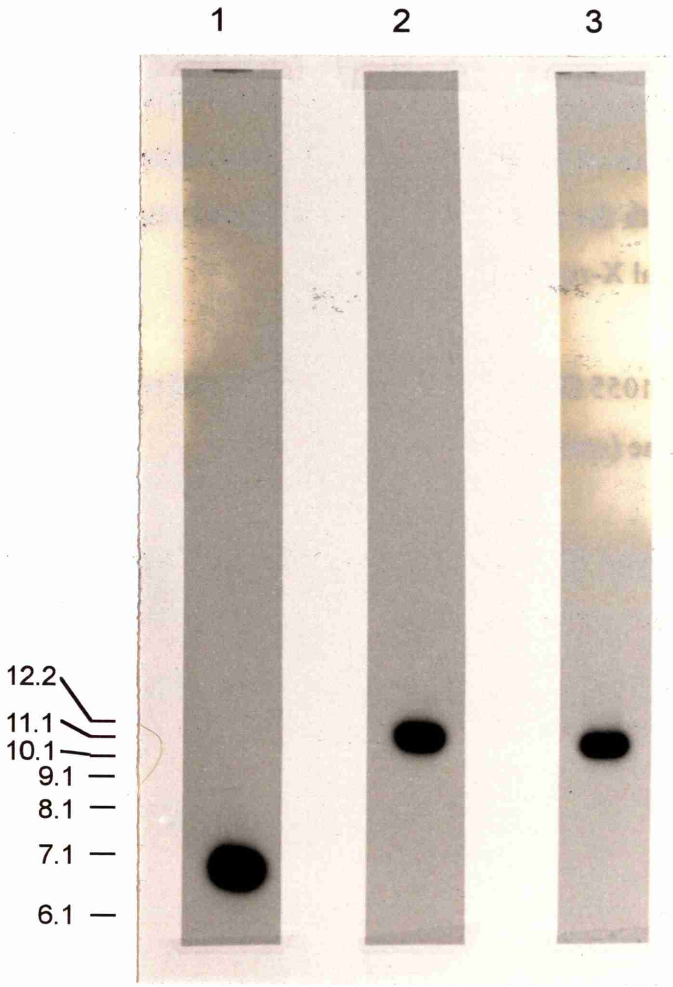

1
2
3
12.2
11.1
10.1
9.1
8.1
7.1
6.1

Figure 3.13

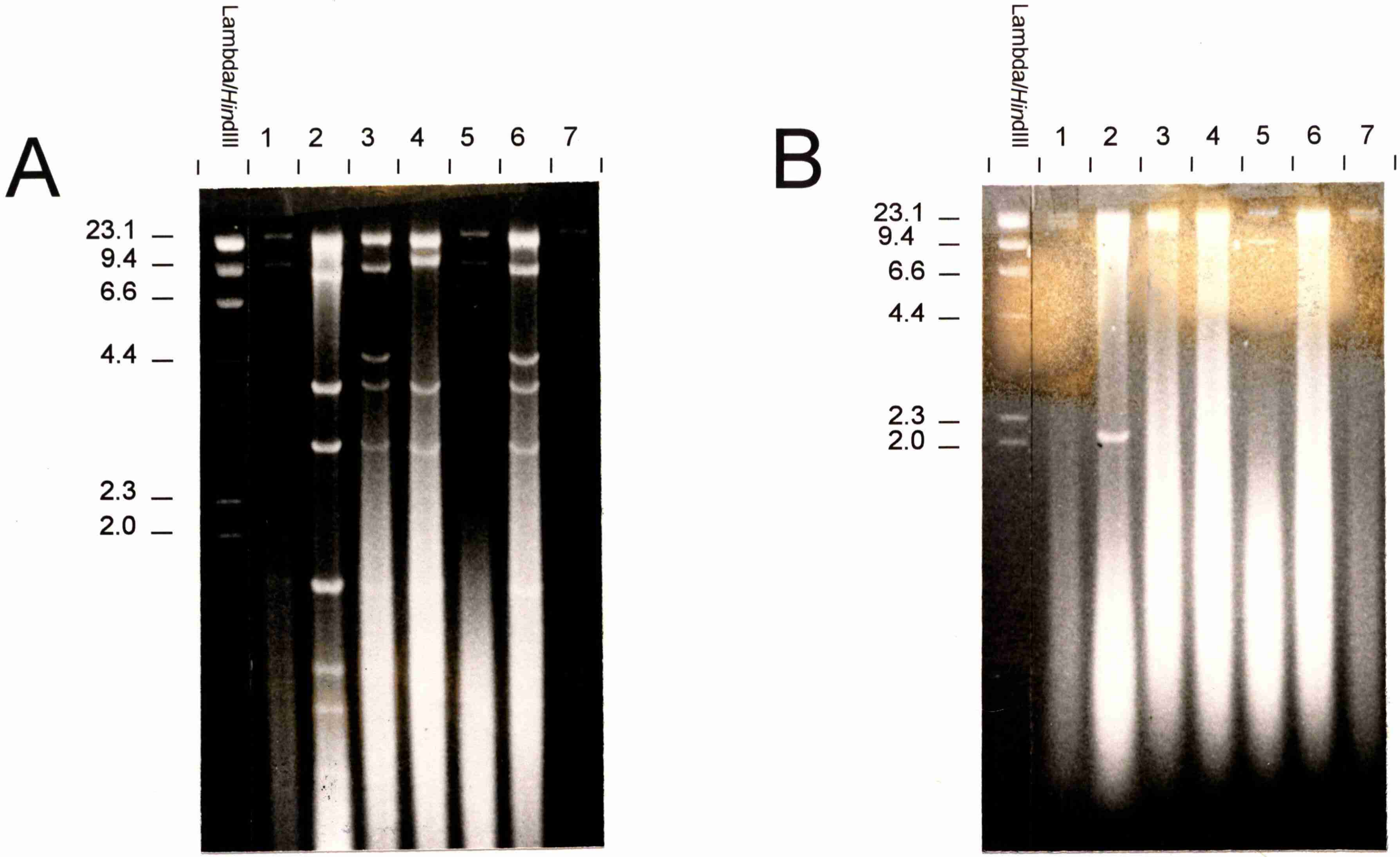
Ethidium bromide stained 0.7% agarose gels of restriction endonuclease digested DNA from the seven positive lambda clones; CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314

A. *Eco*RI digests.

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7
CL#012	CL#034	CL#056	CL#078	CL#0910	CL#01112	CL#01314

B. *Bam*HI digests.

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7
CL#012	CL#034	CL#056	CL#078	CL#0910	CL#01112	CL#01314



A
Lambda/HindIII
1
2
3
4
5
6
7
23.1
9.4
6.6
4.4
2.3
2.0
B
Lambda/HindIII
1
2
3
4
5
6
7
23.1
9.4
6.6
4.4
2.3
2.0

Figure 3.14

Southern blots of restriction endonuclease digested DNA from the seven positive lambda clones; CL#012, CL#034, CL#056, CL#078, CL#0910, CL#01112 and CL#01314

The Southern blots were hybridised with the *cpr-3* cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

A. *Eco*RI digests.

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7
CL#012	CL#034	CL#056	CL#078	CL#0910	CL#01112	CL#01314

B. *Bam*HI digests.

Lane 1	Lane 2	Lane 3	Lane 4	Lane 5	Lane 6	Lane 7
CL#012	CL#034	CL#056	CL#078	CL#0910	CL#01112	CL#01314

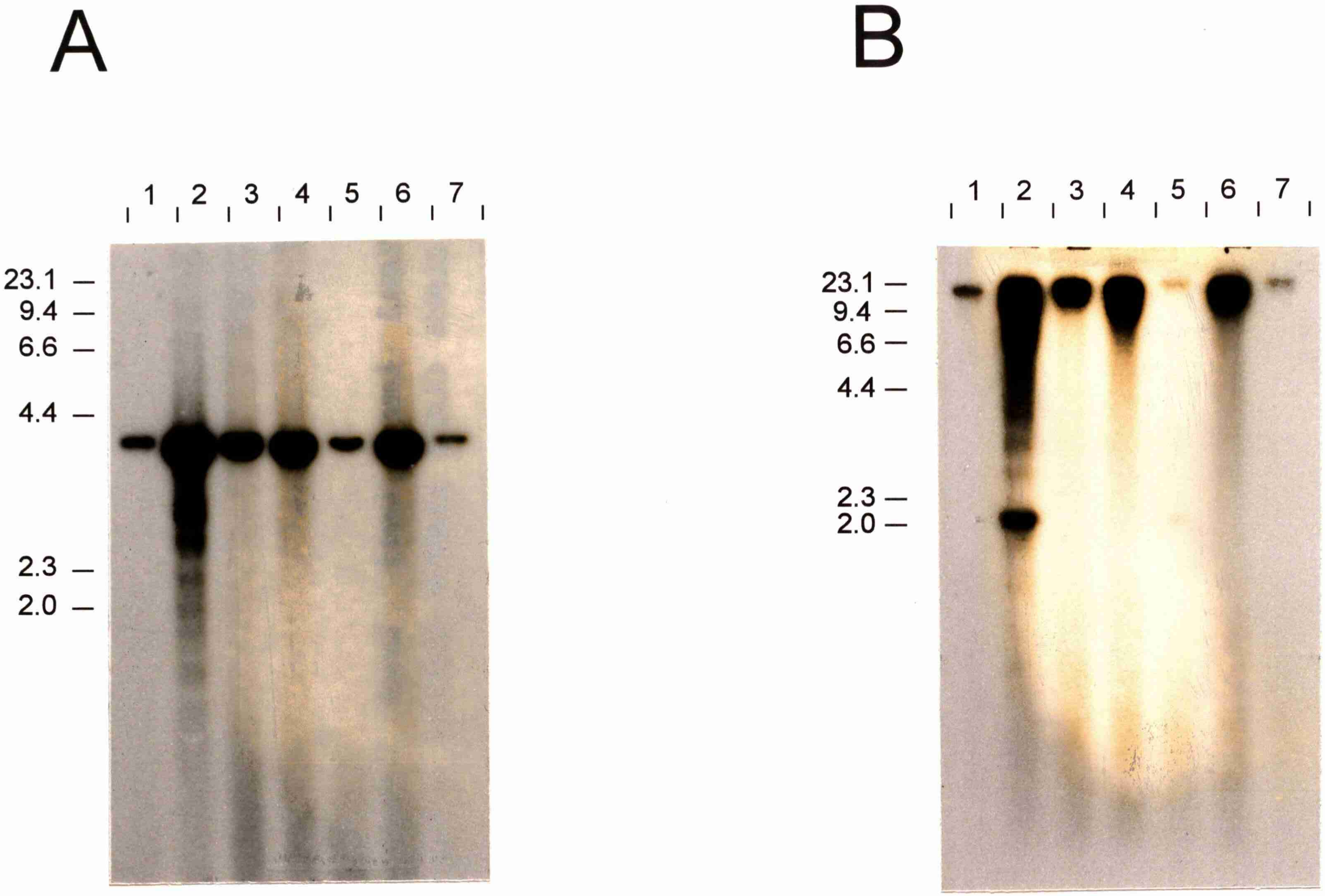

A
B
1
2
3
4
5
6
7
23.1
9.4
6.6
4.4
2.3
2.0

Figure 3.15

Southern blot of CL#034 DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel

The Southern blot was hybridised with the *cpr-3* cDNA clone (cm12b6) and washed at high stringency prior to exposure to medical X-ray film.

Lane 1:	*Pst*I/*Xba*I
Lane 2:	*Pst*I/*Sac*II
Lane 3:	*Pst*I/*Xho*I
Lane 4:	*Xba*I/*Sac*II
Lane 5:	*Xba*I/*Xho*I
Lane 6:	*Xho*I/*Sac*II
Lane 7:	*Apa*I/*Kpn*I

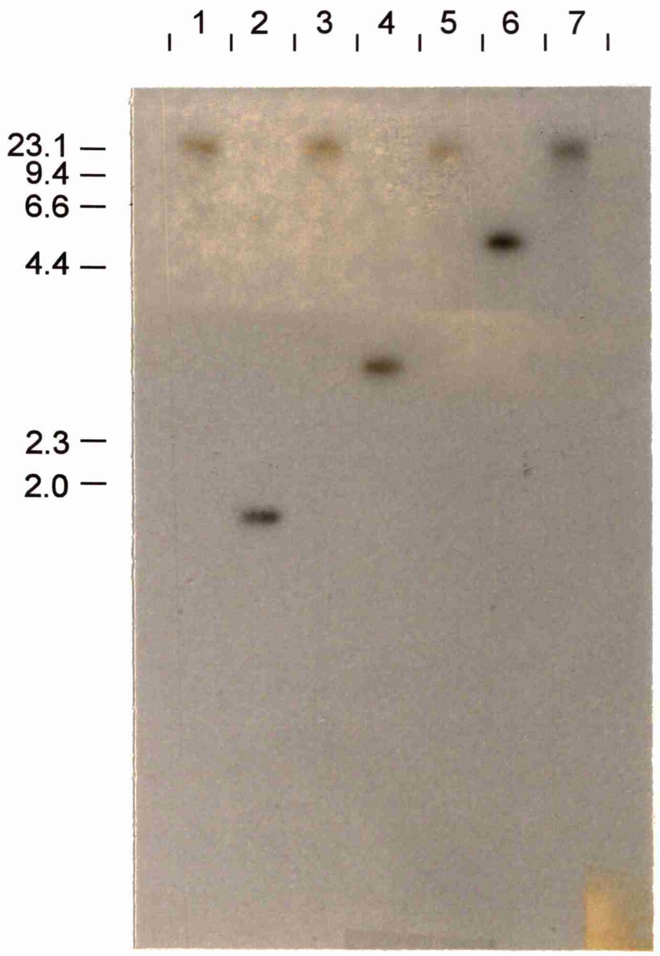

1
2
3
4
5
6
7
23.1
9.4
6.6
4.4
2.3
2.0

Figure 3.16

Restriction map data for CL#034/5/KS+

A. Ethidium bromide stained 0.7% agarose gels of CL#034/5/KS+ DNA digested with a variety of restriction endonucleases

Lane 1: *Hind*III

Lane 2: *Hind*III/*Sac*II

Lane 3: *Hind*III/*Xho*I

Lane 4: *Pst*I/*Sac*II

Lane 5: *Eco*R1/*Sac*II

Lane 6: *Eco*R1/*Xho*I

Lane 7: *Xba*I/*Sac*II

Lane 8: *Xba*I/*Xho*I

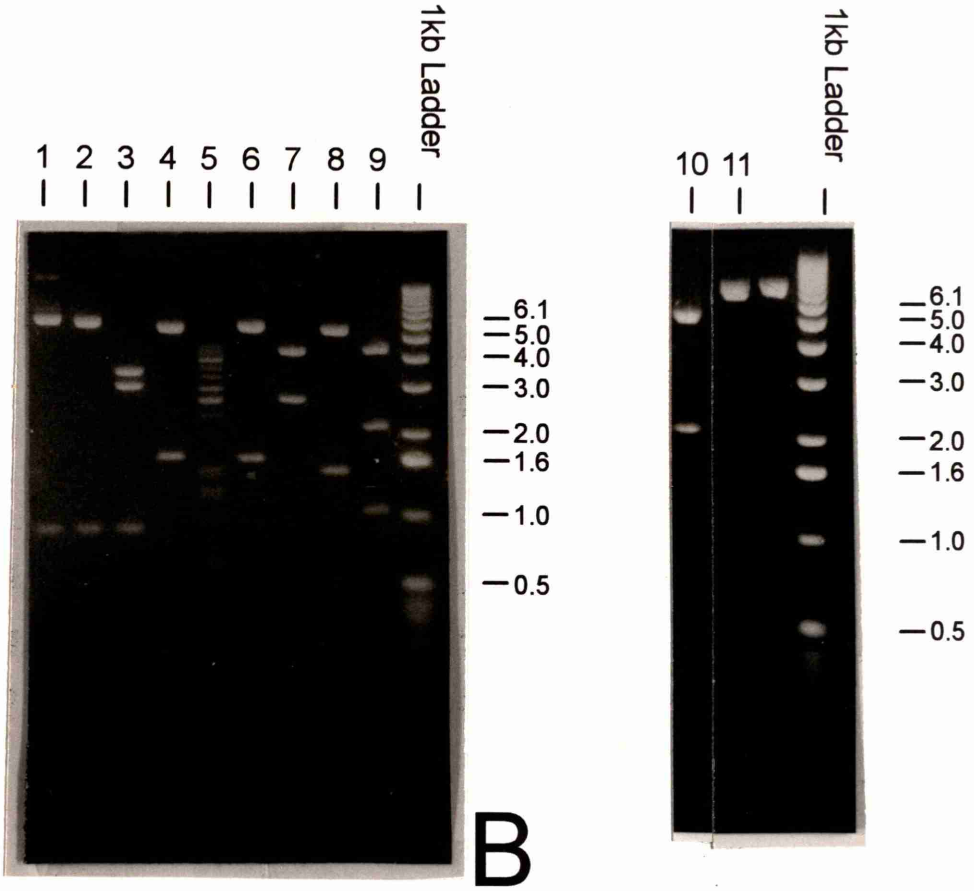
Lane 9: *Cla*I/*Xho*I

Lane 10: *Cla*I

Lane 11: *Xba*I

B. Table summarising the DNA fragment sizes generated by the digests shown in **A**, which were used to generate the restriction map of CL#034/5/KS+. The table also summarises the results obtained after hybridising Southern blots of the gels shown in **A** with the intact cDNA (cm12b6) and the 5′ end cDNA of *cpr-3*.

A



B

Lane Number	Enzymes	Fragment Sizes	Fragment Hybridising to cDNA clone	Fragment Hybridising to 5' end of cDNA clone
Lane 1	*Hind*III	7 kb, 0.9 kb	7 kb, 0.9 kb	7 kb
Lane 2	*Hind*III / *Sac*II	6.8 kb, 0.9 kb, 0.25 kb	6.8 kb, 0.9 kb, 0.25 kb (faint)	0.25 kb
Lane 3	*Hind*III / *Xho*I	3.7 kb, 3.1 kb, 0.9 kb	3.7 kb, 3.1 kb, 0.9 kb	3.1 kb
Lane 4	*Pst*I / *Sac*II	6 kb, 1.7 kb	1.7 kb	1.7 kb
Lane 5	*Eco*RI / *Sac*II	Partial Digest	No Data	No Data
Lane 6	*Eco*RI / *Xho*I	6 kb, 1.7 kb	6 kb	6 kb
Lane 7	*Xba*I / *Sac*II	4.5 kb, 2.7 kb, 0.5 kb	2.7 kb	2.7 kb
Lane 8	*Xba*I / *Xho*I	5.8 kb, 1.5 kb, 0.5 kb	5.8 kb	5.8 kb
Lane 9	*Cla*I / *Xho*I	4.6 kb, 2.2 kb, 1.1 kb	4.5 kb	4.5 kb
Lane 10	*Cla*I	5.8 kb, 2.2 kb	No Data	No Data
Lane 11	*Xba*I	7.5 kb, 0.5 kb	No Data	No Data

Figure 3.17

Restriction map data for ZK1055/6.8/KS-

A. Ethidium bromide stained 0.7% agarose gels of ZK1055/6.8/KS- DNA digested with a variety of restriction endonucleases

Lane 1: *Bam*HI

Lane 2: *Cla*I

Lane 3: *Bam*HI/*Kpn*I

Lane 4: *Kpn*I

Lane 5: *Bam*HI/*Cla*I

Lane 6: *Pst*I

Lane 7: *Pst*I/*Cla*I

Lane 8: *Pst*I/*Bam*HI

Lane 9: *Pst*I/*Kpn*I

Lane 10: *Xba*I

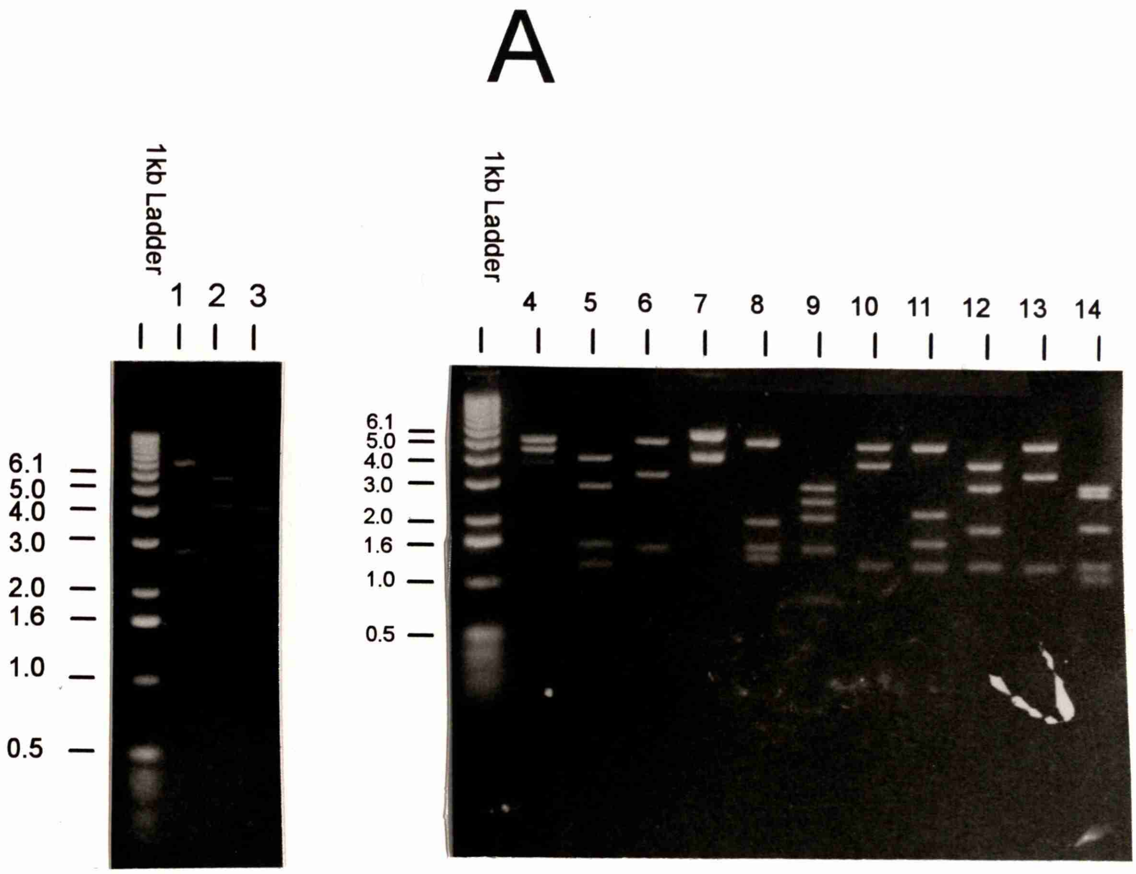
Lane 11: *Xba*I/*Bam*HI

Lane 12: *Xba*I/*Cla*I

Lane 13: *Xba*I/*Pst*I

Lane 14: *Xba*I/*Kpn*I

B. Table summarising the DNA fragment sizes generated by the digests shown in **A**, which were used to generate the restriction map of ZK1055/6.8/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in **A** with the intact cDNA (cm14e3) and the 5′ end cDNA of *cpr-4*.



B

Lane Number	Enzymes	Fragment Sizes	Fragment Hybridising to cDNA clone	Fragment Hybridising to 5' end of cDNA clone
Lane 1	*Bam*HI	7.2 kb, 2.8 kb	7.2 kb, 2.8 kb	2.8 kb
Lane 2	*Cla*I	5.7 kb, 4.2 kb	5.7 kb	5.7 kb
Lane 3	*Bam*HI / *Kpn*I	4.1 kb, 2.9 kb, 2.2 kb, 0.6 kb	4.1 kb, 0.6 kb	0.6 kb
Lane 4	*Kpn*I	5.3 kb, 4.6 kb	4.6 kb	4.6 kb
Lane 5	*Bam*HI / *Cla*I	4.1 kb, 2.9 kb, 1.55 kb, 1.2 kb	4.1 kb, 1.55 kb	1.55 kb
Lane 6	*Pst*I	5.1 kb, 3.3 kb, 1.5 kb	3.3 kb	3.3 kb
Lane 7	*Pst*I / *Cla*I	*Pst*I Digest Failed	No Data	No Data
Lane 8	*Pst*I / *Bam*HI	5.1 kb, 2.0 kb, 1.5 kb, 1.3 kb	2.0 kb, 1.3 kb	1.3 kb
Lane 9	*Pst*I / *Kpn*I	2.9 kb, 2.5 kb, 2.1 kb, 1.5 kb, 0.8 kb	2.5 kb	2.5 kb
Lane 10	*Xba*I	4.9 kb, 3.7 kb, 1.2 kb	3.7 kb	3.7 kb
Lane 11	*Xba*I / *Bam*HI	4.9 kb, 2.1 kb, 1.55 kb, 1.2 kb	2.1 kb, 1.55 kb	1.55 kb
Lane 12	*Xba*I / *Cla*I	3.7 kb, 2.9 kb, 2.0 kb, 1.2 kb	3.7 kb	3.7 kb
Lane 13	*Xba*I / *Pst*I	4.9 kb, 3.3 kb, 1.2 kb, (0.3 kb)*	3.3 kb	3.3 kb
Lane 14	*Xba*I / *Kpn*I	2.9 kb, 2.7 kb, 2.0 kb, 1.2 kb, 1.0 kb	2.7 kb	2.7 kb

* predicted band, not visualised

Figure 3.18

Restriction map data for W02B2/11/KS-

A. Ethidium bromide stained 0.7% agarose gels of W02B2/11/KS- DNA digested with a variety of restriction endonucleases

Lane 1: *Xho*I

Lane 2: *Xho*I/*Pst*I

Lane 3: *Xho*I/*Xba*I

Lane 4: *Cla*I

Lane 5: *Cla*I/*Sac*I

Lane 6: *Bam*HI

Lane 7: *Bam*HI/*Pst*I

Lane 8: *Bam*HI/*Sac*I

Lane 9: *Bam*HI/*Cla*I

Lane 10: *Xho*I/*Sac*I

Lane 11: *Xho*I/*Bam*HI

Lane 12: *Xho*I/*Cla*I

Lane 13: *Xba*I

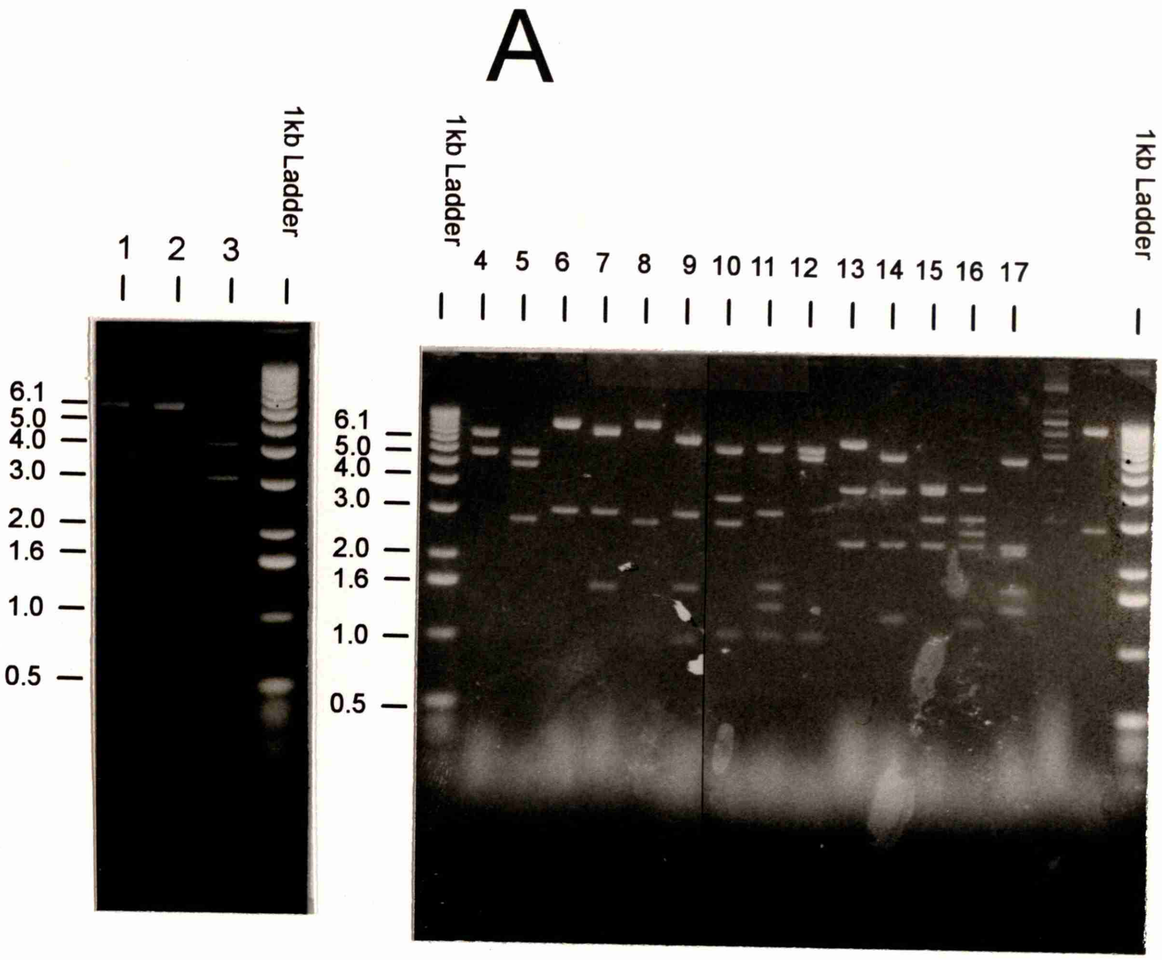
Lane 14: *Xba*I/*Sac*I

Lane 15: *Xba*I/*Pst*I

Lane 16: *Xba*I/*Cla*I

Lane 17: *Xba*I/*Bam*HI

B. Table summarising the DNA fragment sizes generated by the digests shown in **A**, which were used to generate the restriction map of W02B2/11/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in **A** with the intact cDNA (cm04d10) and the 5′ end cDNA of *cpr-5*.



Lane Number	Enzymes	Fragment Sizes	Fragment Hybridising to Intact cDNA	Fragment Hybridising to 5' end of cDNA
Lane 1	*Xho*I	6.2 kb, 6.2 kb, 1.1 kb	6.2 kb	6.2 kb
Lane 2	*Xho*I / *Pst*I	6.2 kb, 6.2 kb, 1.1 kb	6.2 kb	6.2 kb
Lane 3	*Xho*I / *Xba*I	4.2 kb, 3.1 kb, 3.1 kb, 2.2 kb, 0.9 kb, (0.3 kb)*	3.1 kb	3.1 kb
Lane 4	*Cla*I	8 kb, 5.7 kb	5.7 kb	5.7 kb
Lane 5	*Cla*I / *Sac*I	5.7 kb, 4.9 kb, 2.9 kb	5.7 kb	5.7 kb
Lane 6	*Bam*HI	10.5 kb, 3.0 kb	10.5 kb, 3.0 kb	3.0 kb
Lane 7	*Bam*HI/ *Pst*I	9 kb, 3.0 kb, 1.6 kb	3.0 kb, 1.6 kb	3.0 kb
Lane 8	*Bam*HI / *Sac*I	Digest Failed	No data	No data
Lane 9	*Bam*HI / *Cla*I	7.8 kb, 3.0 kb, 1.6 kb, 1.0 kb	3.0 kb, 1.6 kb	3.0 kb
Lane 10	*Xho*I / *Sac*I	6.2 kb, 3.6 kb, 2.9 kb, 1.1 kb	6.2 kb	6.2 kb
Lane 11	*Xho*I / *Bam*HI	6.2 kb, 3.0 kb, 1.6 kb, 1.4 kb, 1.1 kb	3 kb, 1.6 kb	3.0 kb
Lane 12	*Xho*I / *Cla*I	6.2 kb,5.8 kb, 1.1 kb, (0.4 kb)*	5.8 kb	5.8 kb
Lane 13	*Xba*I	7.2 kb, 4 kb, 2.5 kb	7.2 kb	7.2 kb
Lane 14	*Xba*I / *Sac*I	5.8 kb, 4 kb, 2.5 kb, 1.3 kb	5.8 kb	5.8 kb
Lane 15	*Xba*I / *Pst*I	4.2 kb, 4 kb, 3 kb, 2.5 kb	3 kb	3 kb
Lane 16	*Xba*I / *Cla*I	4.2 kb, 3 kb, 2.7 kb, 2.5 kb, 1.25 kb	3 kb	3 kb
Lane 17	*Xba*I / *Bam*HI	5.9 kb, 2.5 kb, 2.4 kb, 1.65 kb, 1.4 kb	5.9 kb, 1.4 kb	1.4 kb

* denote predicted bands (could not visualise on gel)

Figure 3.19

Restriction map data for C25B8/10.7/KS-

A. Ethidium bromide stained 0.7% agarose gels of C25B8/10.7/KS- DNA digested with a variety of restriction endonucleases

Lane 1: *Kpn*I

Lane 2: *Kpn*I/*Pst*I

Lane 3: *Kpn*I/*Sac*I

Lane 4: *Sal*I

Lane 5: *Sal*I/*Sac*I

Lane 6: *Sal*I/*Kpn*I

Lanes 7 and 11: *Bam*HI

Lane 8: *Bam*HI/*Kpn*I

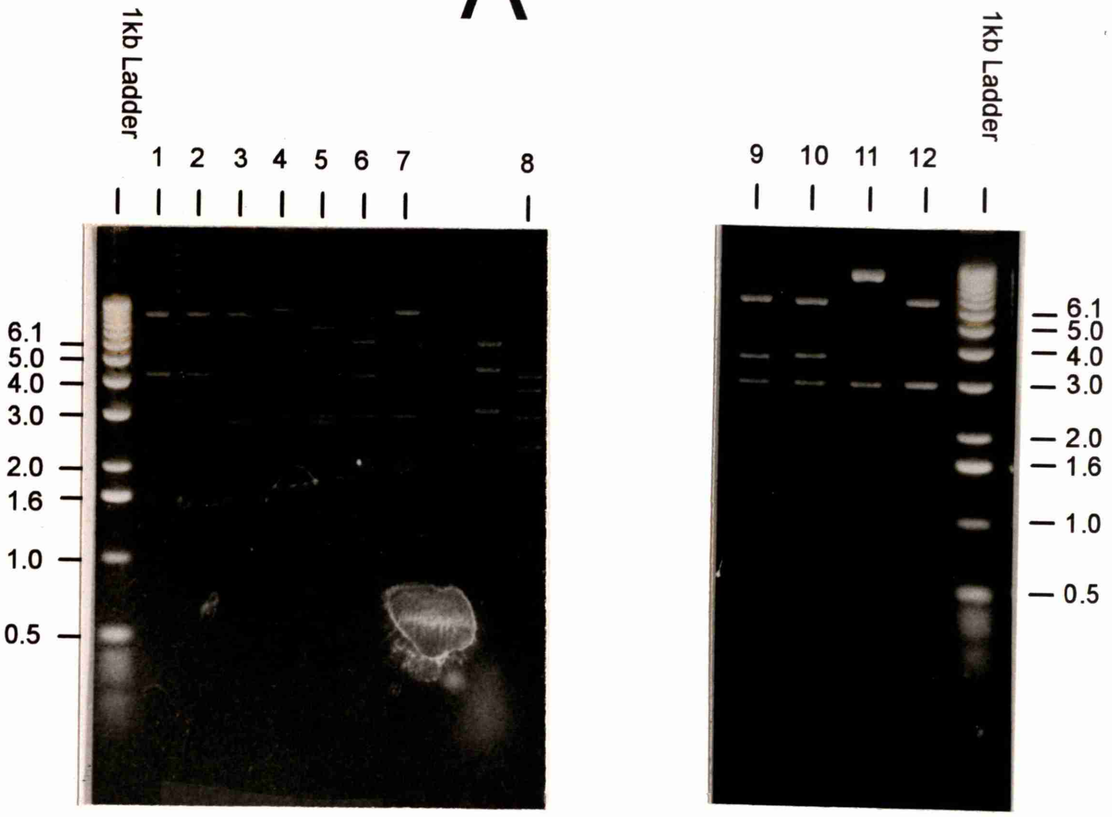
Lane 9: *Bam*HI/*Pst*I

Lane 10: BamHI/*Sac*I

Lane 12: BamHI/*Sal*I

B. Table summarising the DNA fragment sizes generated by the digests shown in **A**, which were used to generate the restriction map of C25B8/10.7/KS-. The table also summarises the results obtained after hybridising Southern blots of the gels shown in **A** with the intact cDNA (cm01a5) and the 5′ end cDNA of *cpr-6.*

A



B

Lane Number	Enzymes	Fragment Sizes	Fragment Hybridising to intact cDNA clone	Fragment Hybridising to 5' end of cDNA clone
Lane 1	*Kpn*I	9.5 kb, 4.2 kb	9.5 kb	9.5 kb
Lane 2	*Kpn*I / *Pst*I	9.5 kb, 4.2 kb	9.5 kb	9.5 kb
Lane 3	*Kpn*I / *Sac*I	9.5 kb, 2.9 kb, 1.4 kb	9.5 kb	9.5 kb
Lane 4	*Sal*I	11 kb, 3 kb	11 kb	11 kb
Lane 5	*Sal*I / *Sac*I	7.8 kb, 3 kb, 2.9 kb	7.8 kb	7.8 kb
Lane 6	*Sal*I / *Kpn*I	6.5 kb, 4.2 kb, 3 kb	6.5 kb	6.5 kb
Lane 7 and 11	*Bam*HI	11 kb, 3 kb	11 kb (faint),3 kb	11 kb
Lane 8	*Bam*HI / *Kpn*I	4.2 kb, 3.8 kb, 3 kb, 2.4 kb	3 kb, 3.8 kb (faint)	3.8 kb
Lane 9	*Bam*HI / *Pst*I	7 kb, 3.8 kb, 3 kb	3.8 kb, 3 kb	3.8 kb
Lane 10	*Bam*HI / *Sac*I	6.9 kb, 3.9 kb, 3 kb	6.9 kb, 3 kb	6.9 kb
Lane 12	*Bam*HI / *Sal*I	7 kb, 3 kb, 3 kb, 0.7 kb	3 kb, 0.7 kb	0.7 kb

Figure 3.20

The restriction map of CL#034/5/KS+

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5′ - 3′) and maximum size of the *cpr-3* gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of CL#034/5/KS+ (shown in Figure 3.16A and 3.16B) are also summarised below the map.

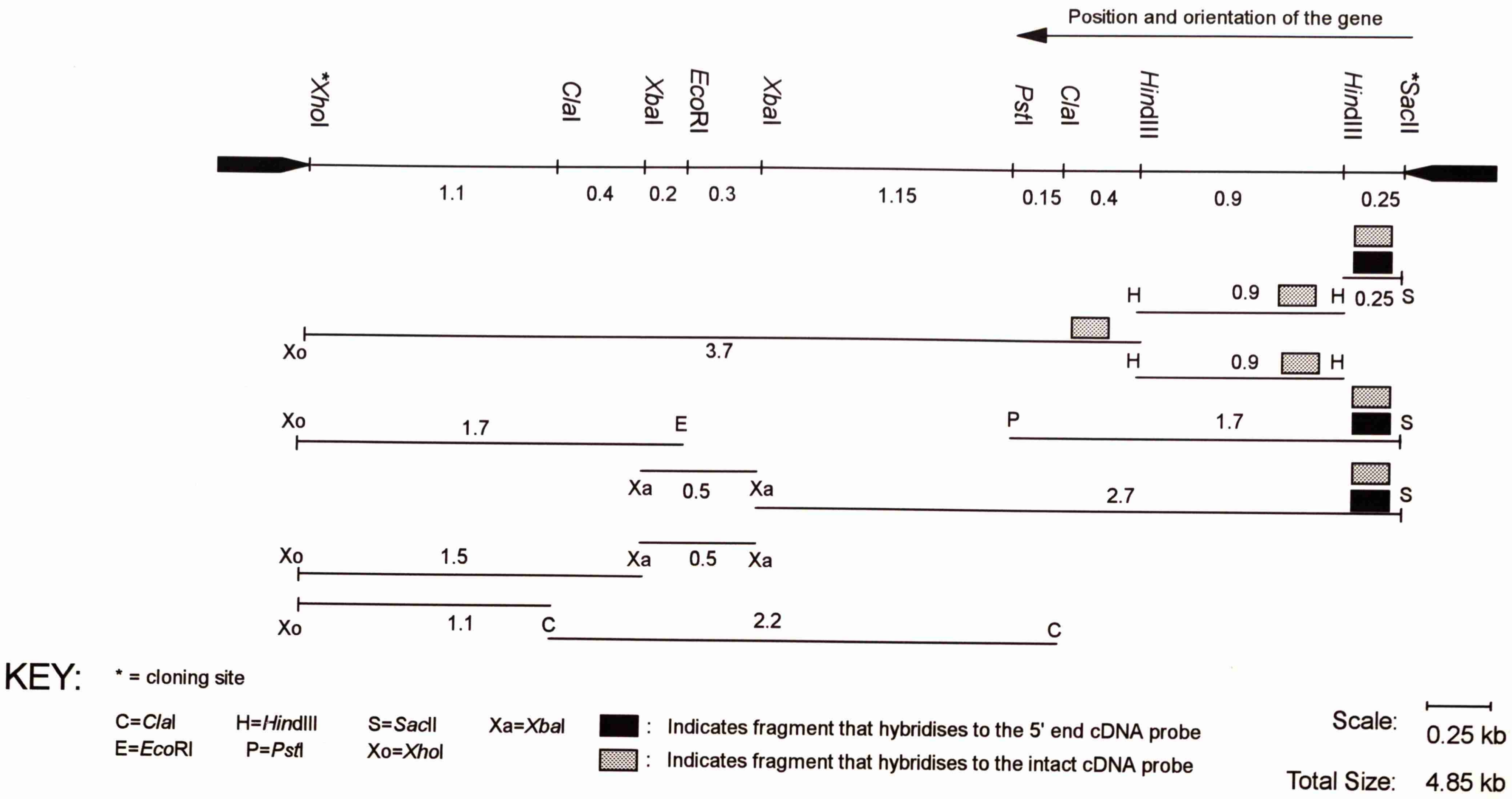
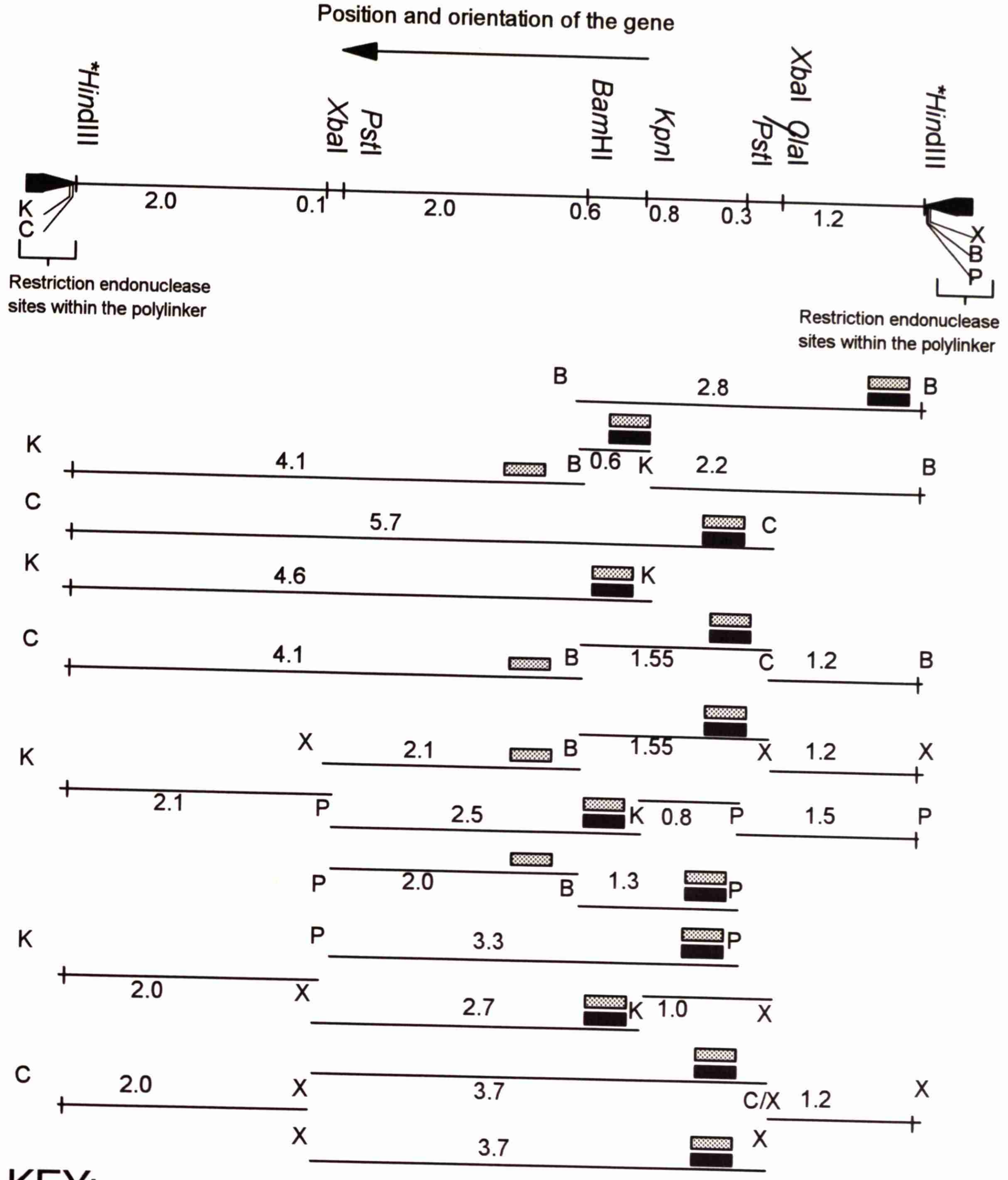

Position and orientation of the gene
*XhoI
ClaI
XbaI
EcoRI
XbaI
PstI
ClaI
HindIII
HindIII
*SacII
1.1
0.4
0.2
0.3
1.15
0.15
0.4
0.9
0.25
H
0.9
H
0.25
S
Xo
3.7
H
0.9
H
Xo
1.7
E
P
1.7
S
Xa
0.5
Xa
2.7
S
Xo
1.5
Xa
0.5
Xa
Xo
1.1
C
2.2
C
KEY:
* = cloning site
C=ClaI
E=EcoRI
H=HindIII
P=PstI
S=SacII
Xo=XhoI
Xa=XbaI
: Indicates fragment that hybridises to the 5' end cDNA probe
: Indicates fragment that hybridises to the intact cDNA probe
Scale:
0.25 kb
Total Size:
4.85 kb

Figure 3.21

The restriction map of ZK1055/6.8/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5′ - 3′) and maximum size of the *cpr-4* gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of ZK1055/6.8/KS- (shown in Figure 3.17A and 3.17B) are also summarised below the map.

Note: The position of the *Hin*dIII cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- is indicated by a vertical line on the appropriate fragments.



KEY:

B=*BamH*I
C=*Cla*I
K=*Kpn*I
P=*Pst*I
X=*Xba*I

* = cloning site

■ : Indicates fragment that hybridises to the 5' end cDNA probe
▨ : Indicates fragment that hybridises to the intact cDNA probe

Total Size: 7 kb

Scale: 0.5 kb

Figure 3.22

The restriction map of W02B2/11/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5′ - 3′) and maximum size of the *cpr-5* gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of W02B2/11/KS- (shown in Figure 3.18A and 3.18B) are also summarised below the map.

Note: The position of the *Pst*I and *Sac*I cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- are indicated by a vertical line on the appropriate fragments.

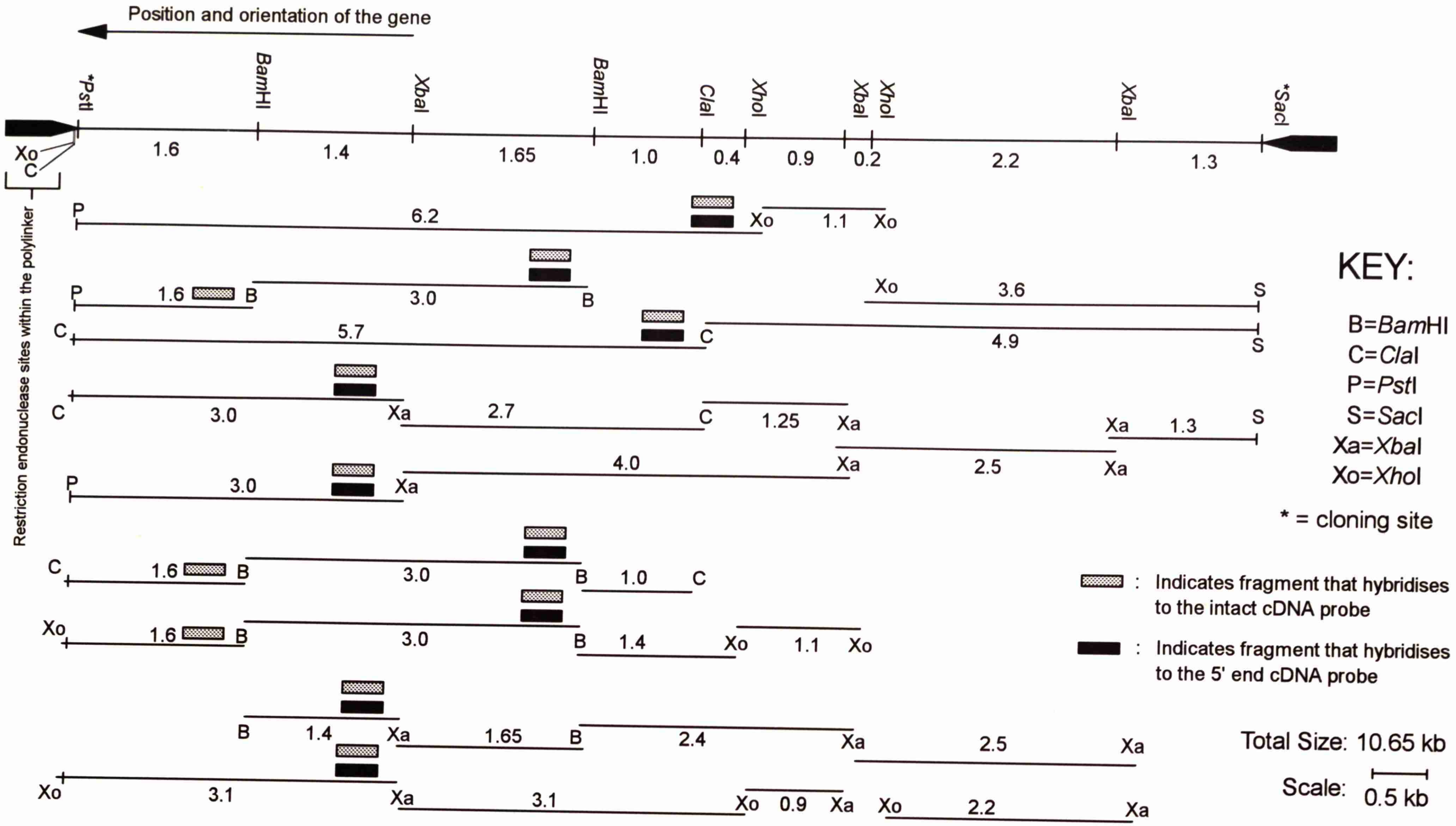

Position and orientation of the gene
*PstI
BamHI
XbaI
BamHI
ClaI
XhoI
XbaI
XhoI
XbaI
*SacI
Xo
C
Restriction endonuclease sites within the polylinker
1.6
1.4
1.65
1.0
0.4
0.9
0.2
2.2
1.3
P
6.2
Xo
1.1
Xo
P
1.6
B
3.0
B
Xo
3.6
S
C
5.7
C
4.9
S
C
3.0
Xa
2.7
C
1.25
Xa
Xa
1.3
S
P
3.0
Xa
4.0
Xa
2.5
Xa
C
1.6
B
3.0
B
1.0
C
Xo
1.6
B
3.0
B
1.4
Xo
1.1
Xo
B
1.4
Xa
1.65
B
2.4
Xa
2.5
Xa
Xo
3.1
Xa
3.1
Xo
0.9
Xa
Xo
2.2
Xa
KEY:
B=BamHI
C=ClaI
P=PstI
S=SacI
Xa=XbaI
Xo=XhoI
* = cloning site
▨ : Indicates fragment that hybridises to the intact cDNA probe
■ : Indicates fragment that hybridises to the 5' end cDNA probe
Total Size: 10.65 kb
Scale: 0.5 kb

Figure 3.23

The restriction map of C25B8/10.7/KS-

The restriction map shows the position of each mapped restriction endonuclease site and the distances between these sites (in kb). The map also shows the position, orientation (5′ - 3′) and maximum size of the *cpr-6* gene within the subclone (marked by the arrow), obtained from the hybridisation data. The data used to generate the restriction map of C25B8/10.7/KS- (shown in Figure 3.19A and 3.19B) are also summarised below the map.

Note: The position of the *Pst*I and *Sac*I cloning site with respect to the restriction endonuclease sites in the polylinker of pBluescript II KS- are indicated by a vertical line on the appropriate fragments.

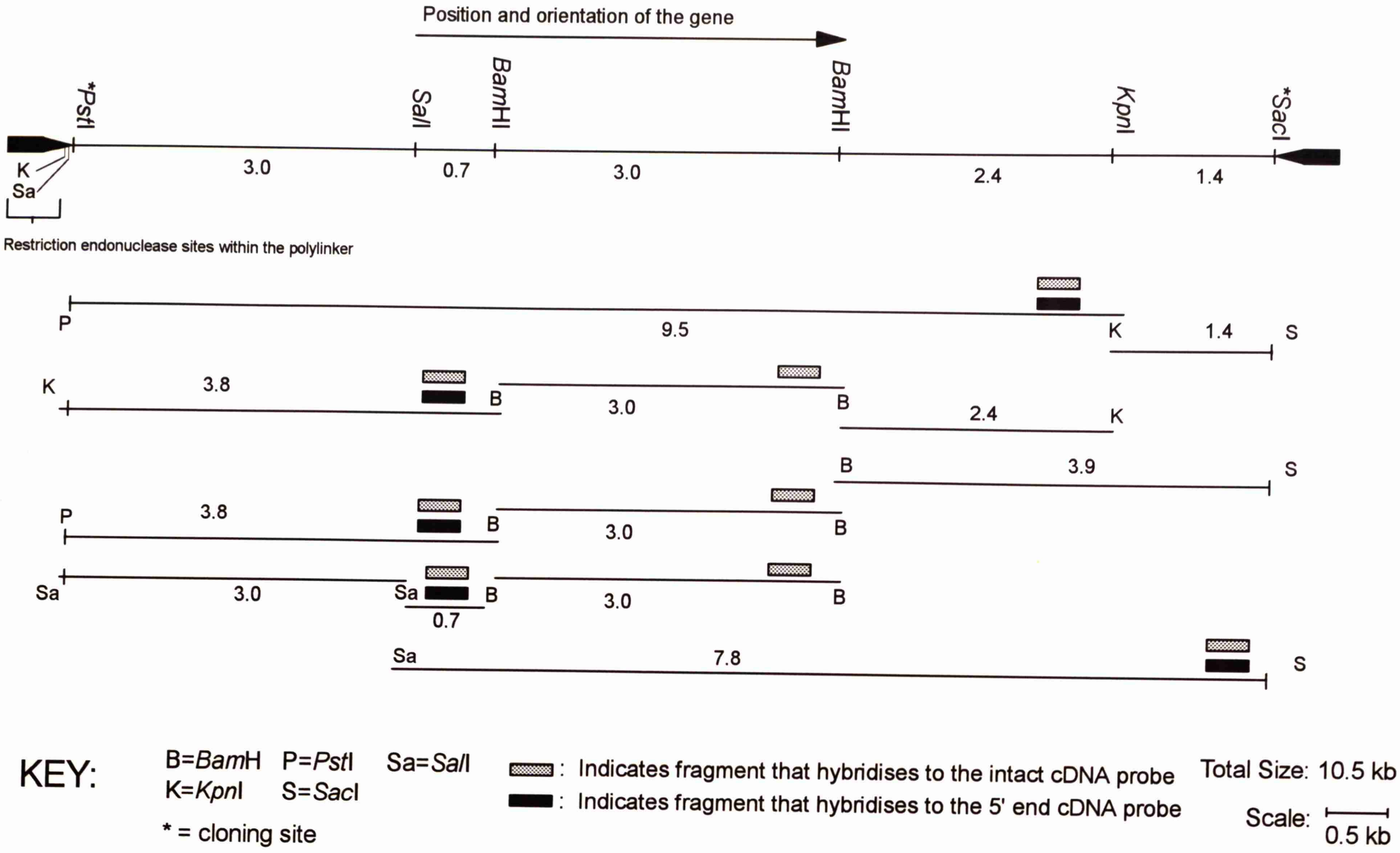

Position and orientation of the gene
*PstI
SalI
BamHI
BamHI
KpnI
*SacI
3.0
0.7
3.0
2.4
1.4
K
Sa
Restriction endonuclease sites within the polylinker
P
9.5
K
1.4
S
K
3.8
B
3.0
B
2.4
K
B
3.9
S
P
3.8
B
3.0
B
Sa
3.0
Sa
0.7
B
3.0
B
Sa
7.8
S
KEY:
B=BamH P=PstI Sa=SalI
K=KpnI S=SacI
* = cloning site
▨ : Indicates fragment that hybridises to the intact cDNA probe
■ : Indicates fragment that hybridises to the 5' end cDNA probe
Total Size: 10.5 kb
Scale: 0.5 kb

Figure 3.24

A summary of the nucleotide sequence identity shared between *cpr-3* and its cDNA clone, cm12b6, upstream of the *Sac*II site used for cloning the insert of CL#034/5/KS+

```
                                 gagttttcagttttccctcaaatttca
-80 tgtgcggtgtgtgtttgctgaaacgcttctttccagttttcagttttccctcaaatttca -21

    aaaattgaatactaaagagaATGCTGAAAGTGTACTTTTTG
-20 aaaattgaatactaaagagaATGCTGAAAGTGTACTTTTTGgtgagtcatcttcatcttc 40
                         M  L  K  V  Y  F  L

                                 GCACTGTTTCTAGCCGGGTGCTCTGCATTTG
 41 atttttccattatttttcttacatttagGCACTGTTTCTAGCCGGGTGCTCTGCATTTG 100
                                 A  L  F  L  A  G  C  S  A  F  V
    TTCTTGATGAAATCCGCGG
101 TTCTTGATGAAATCCGCGG
                 SacII
```

KEY:

Bold:	Nucleotide sequence of the 5' end of the *cpr-3* cDNA clone (cm12b6) up to and including the *Sac*II site.
Underlined:	Nucleotide sequence of the 5' end of the *cpr-3* gene up to and including the *Sac*II site.
Uppercase:	Coding sequence, the amino acid residue encoded by each codon is given below the appropriate codon.
Lowercase:	Non-coding sequence

Figure 3.25

A summary of the restriction map and DNA sequence data obtained from the subclones, CL#034/5/KS+ and CL#034/1.7/KS+

The figure shows the position and size of the two subclones, CL#034/5/KS+ and CL#034/1.7/KS+ with respect to one another. The position of the restriction endonuclease sites determined for each of these subclones by restriction mapping are also shown. A summary of the DNA sequence of the *cpr-3* gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5′-3′, marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of *C.elegans* genomic DNA probed with the *cpr-3* cDNA (cm12b6). The extent of the region covered by the subclones which is known to be contiguous with the *C.elegans* genome is also indicated.

NOTES:

***Eco*RV**: The *Eco*RV site is marked by an asterisk because its position was not determined from the original restriction mapping experiments. This site was originally identified from the DNA sequence data. Subsequently the presence and position of this single site within CL#034/1.7/KS+ was confirmed by double digests using *Hind*III and *Eco*RV. Later triple digests of CL#034/5/KS+ with *Eco*RV, *Hind*III and *Sac*II demonstrated that no *Eco*RV site was present in this subclone.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale (shown in the figure). The DNA sequence of the *cpr-3* gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.

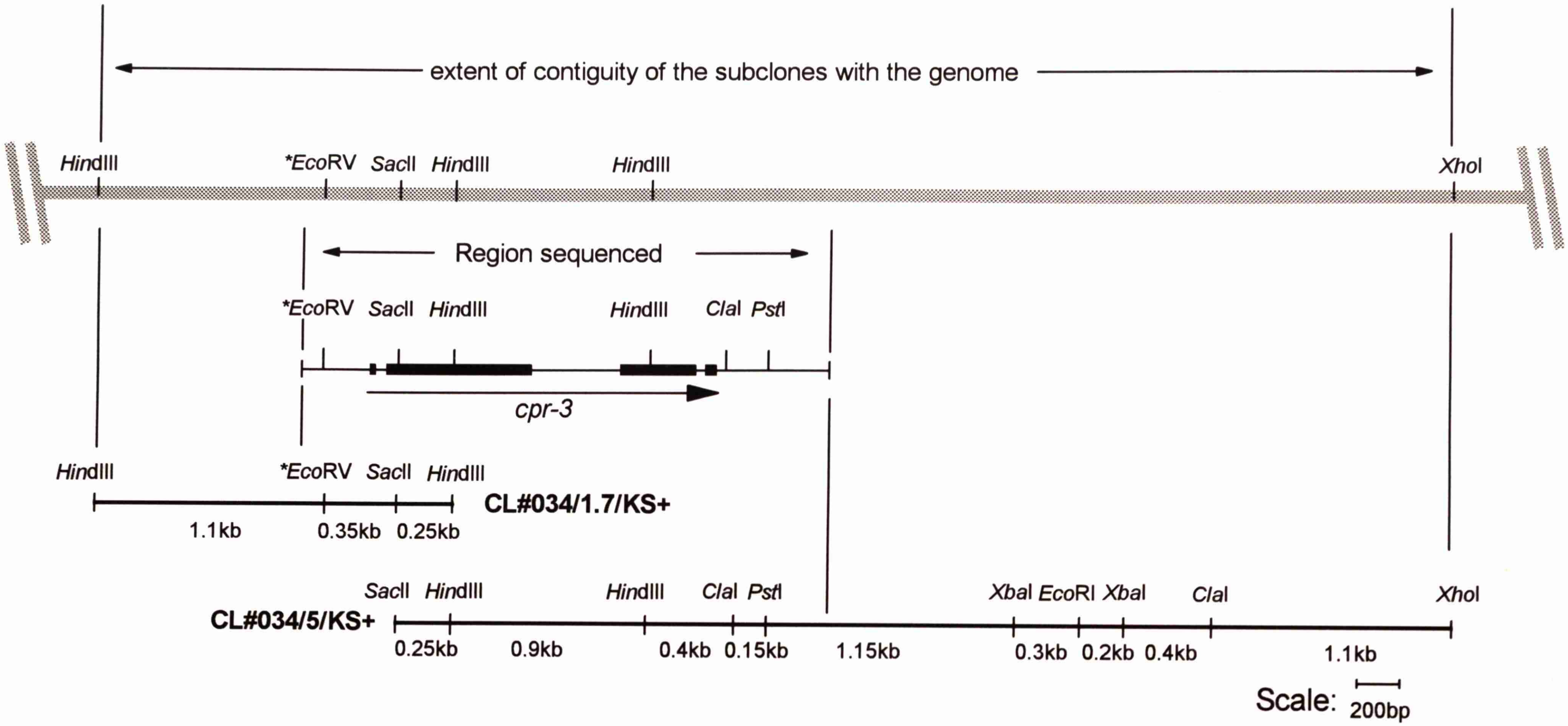

extent of contiguity of the subclones with the genome
HindIII
*EcoRV
SacII
HindIII
HindIII
XhoI
Region sequenced
*EcoRV
SacII
HindIII
HindIII
ClaI
PstI
cpr-3
HindIII
*EcoRV
SacII
HindIII
CL#034/1.7/KS+
1.1kb
0.35kb
0.25kb
SacII
HindIII
HindIII
ClaI
PstI
XbaI
EcoRI
XbaI
ClaI
XhoI
CL#034/5/KS+
0.25kb
0.9kb
0.4kb
0.15kb
1.15kb
0.3kb
0.2kb
0.4kb
1.1kb
Scale:
200bp

Figure 3.26

A summary of the restriction map and DNA sequence data obtained from the subclones, ZK1055/6.8/KS- and ZK1055/2.8/KS+

The figure shows the position and size of the two subclones ZK1055/6.8/KS- and ZK1055/2.8/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these subclones. A summary of the DNA sequence of the *cpr-4* gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5′-3′, marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of *C.elegans* genomic DNA probed with the *cpr-4* cDNA (cm14e3). The extent of the region covered by the subclones which is known to be contiguous with the *C.elegans* genome is also indicated.

NOTES: The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the *cpr-4* gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.

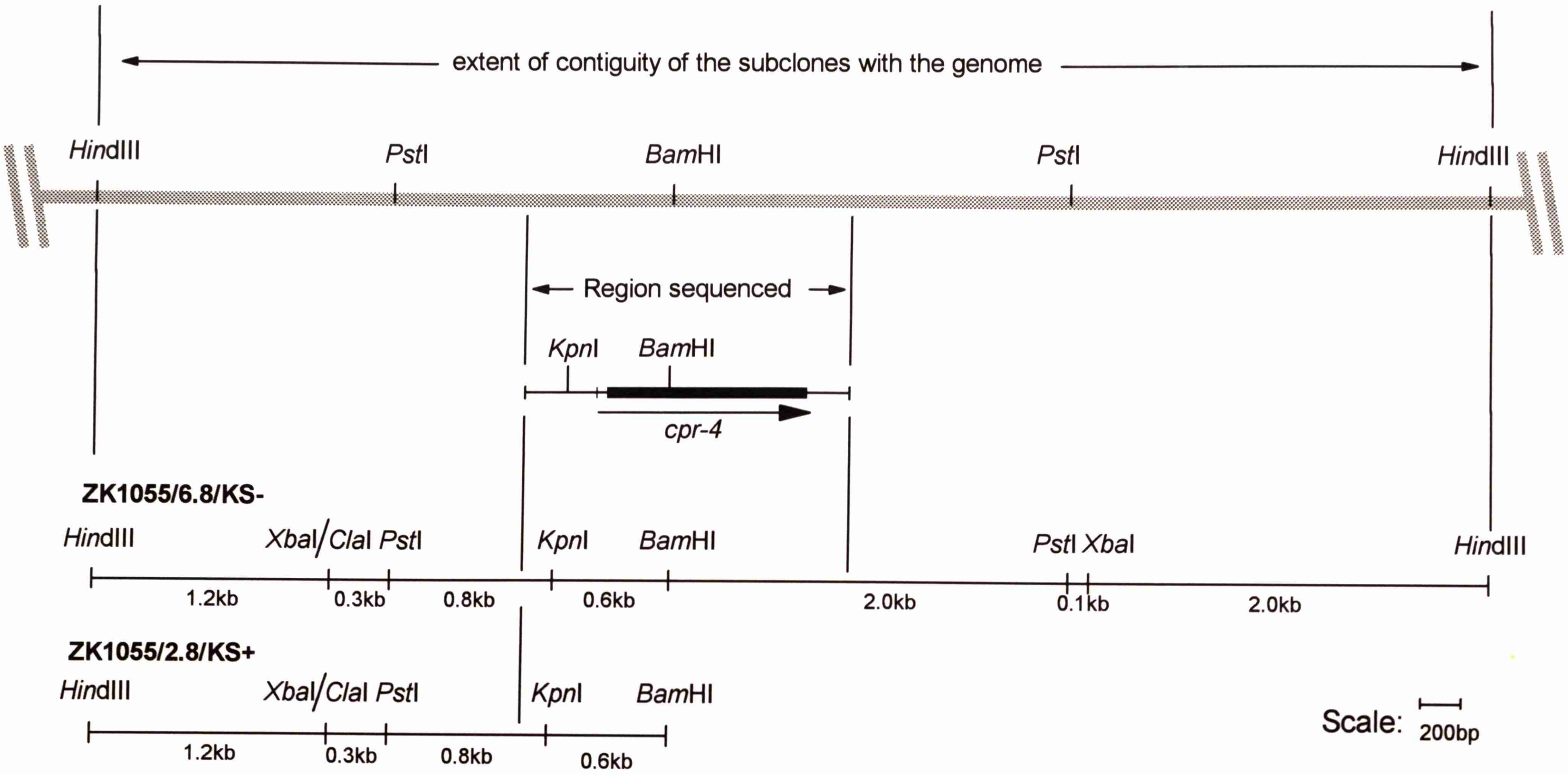

extent of contiguity of the subclones with the genome
HindIII
PstI
BamHI
PstI
HindIII
Region sequenced
KpnI
BamHI
cpr-4
ZK1055/6.8/KS-
HindIII
XbaI/ClaI PstI
KpnI
BamHI
PstI XbaI
HindIII
1.2kb
0.3kb
0.8kb
0.6kb
2.0kb
0.1kb
2.0kb
ZK1055/2.8/KS+
HindIII
XbaI/ClaI PstI
KpnI
BamHI
1.2kb
0.3kb
0.8kb
0.6kb
Scale:
200bp

Figure 3.27

A summary of the restriction map and DNA sequence data obtained from the subclones, W02B2/11/KS-, W02B2/3/KS+ and W02B2/1.6/KS+

The figure shows the position and size of the three subclones W02B2/11/KS-, W02B2/3/KS+ and W02B2/1.6/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these subclones. A summary of the DNA sequence of the *cpr-5* gene, obtained from the two subclones, is shown above the subclones. This summary shows the orientation of the gene (5′-3′, marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of *C.elegans* genomic DNA probed with the *cpr-5* cDNA (cm04d10). The extent of the region covered by the subclones which is known to be contiguous with the *C.elegans* genome is also indicated.

NOTES:

***BssHII*:** The *Bss*HII site marked with the asterisk was originally identified from the DNA sequence data. The presence and position of a single *Bss*HII site in W02B2/3/KS+ site was subsequently confirmed by *Bss*HII digests of this subclone. It is not known whether *Bss*HII sites are also present in the region covered by the cloned insert of W02B2/1.6/KS+ since this was not tested.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the *cpr-5* gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.

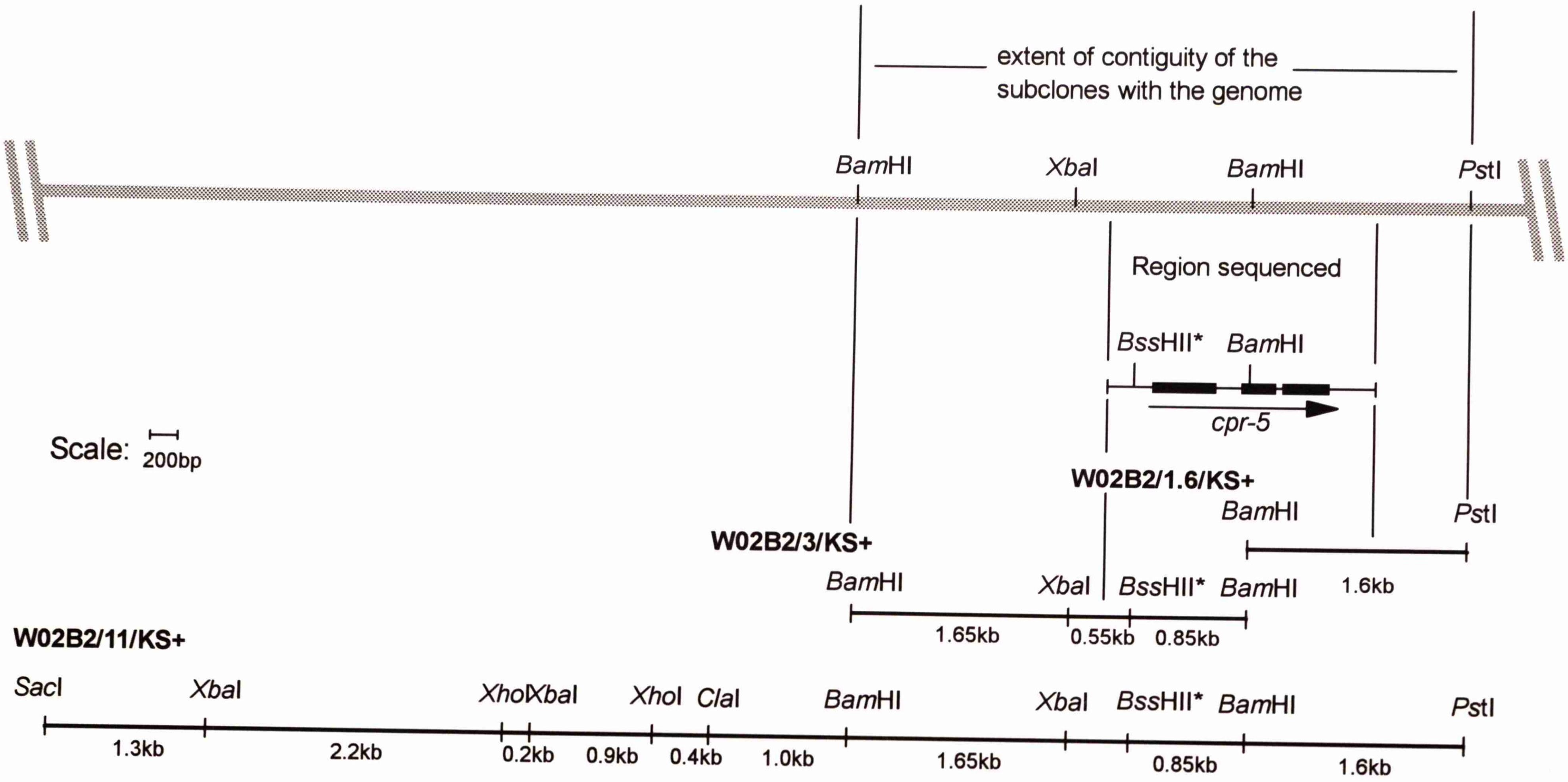

extent of contiguity of the
subclones with the genome
BamHI
XbaI
BamHI
PstI
Region sequenced
BssHII*
BamHI
cpr-5
Scale:
200bp
W02B2/1.6/KS+
BamHI
PstI
1.6kb
W02B2/3/KS+
BamHI
XbaI
BssHII*
BamHI
1.65kb
0.55kb
0.85kb
W02B2/11/KS+
SacI
XbaI
XhoI
XbaI
XhoI
ClaI
BamHI
XbaI
BssHII*
BamHI
PstI
1.3kb
2.2kb
0.2kb
0.9kb
0.4kb
1.0kb
1.65kb
0.85kb
1.6kb

Figure 3.28

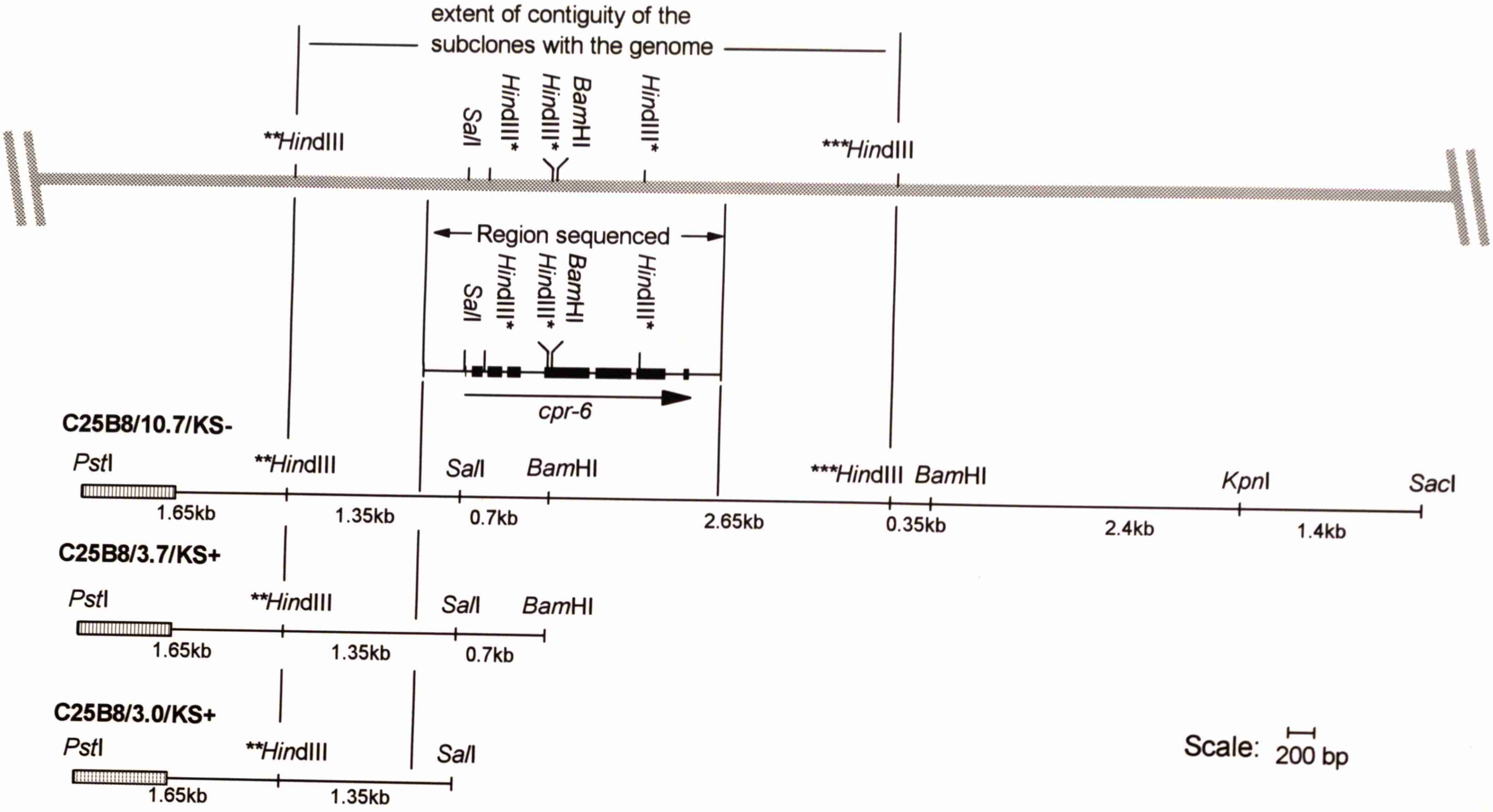
A summary of the restriction map and DNA sequence data obtained from the subclones, C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3/KS+

The figure shows the position and size of the three subclones C25B8/10.7/KS-, C25B8/3.7/KS+ and C25B8/3/KS+ with respect to one another. The position of the restriction endonuclease sites determined by restriction mapping are also shown for each of these clones. The striped boxes (▥) indicate the presence of approximately 700 bp of pJB8 cosmid vector sequence. A summary of the DNA sequence of the *cpr-6* gene, obtained from the three subclones, is shown above the subclones. This summary shows the orientation of the gene (5′-3′, marked by the arrow), the coding regions of the gene (marked by black boxes), the introns (marked by the thin lines between boxes) and the mapped restriction endonuclease sites that were confirmed by sequencing, with the exception of the *Hind*III sites which will be discussed below. Above the DNA sequence summary, the thick line represents the region of the genome covered by the subclones. The marked restriction endonuclease sites represent the sites known to be contiguous with the subclones, determined from Southern blots of diagnostic digests of *C.elegans* genomic DNA probed with the *cpr-6* cDNA (cm01a5). The extent of the region covered by the subclones which is known to be contiguous with the *C.elegans* genome is also indicated.

NOTES:

***Hind*III sites:** The *Hind*III sites marked by a single asterisk were originally identified from the DNA sequence of the *cpr-6* gene. The position of the *Hind*III site marked by two asterisks in C25B8/3.7/KS+ was subsequently determined by restriction mapping using single (*Hind*III) and double (*Hind*III/*Sal*I) digests. The *Hind*III site marked by three asterisks was determined from Southern blots of diagnostic digests of both *C.elegans* genomic DNA and C25B8/10.7/KS- DNA probed with the *cpr-6* cDNA (cm01a5) which revealed there was a previously undetermined *Hind*III site 2.0 kb downstream of the most 3′ *Hind*III site within the *cpr-6* gene.

The size of the subclones, the position of the restriction maps, the extent of the sequenced region and the sizes of the coding regions and introns are all drawn to the same scale, shown in the figure. The DNA sequence of the *cpr-6* gene was aligned with the subclones, by aligning those restriction endonuclease sites identified by both restriction mapping and DNA sequencing.



extent of contiguity of the subclones with the genome
**HindIII
SalI
HindIII*
HindIII*
BamHI
HindIII*
***HindIII
Region sequenced
SalI
HindIII*
HindIII*
BamHI
HindIII*
cpr-6
C25B8/10.7/KS-
PstI
**HindIII
SalI
BamHI
***HindIII
BamHI
KpnI
SacI
1.65kb
1.35kb
0.7kb
2.65kb
0.35kb
2.4kb
1.4kb
C25B8/3.7/KS+
PstI
**HindIII
SalI
BamHI
1.65kb
1.35kb
0.7kb
C25B8/3.0/KS+
PstI
**HindIII
SalI
1.65kb
1.35kb
Scale: 200 bp

Figure 3.29

A Southern blot of CL#034 DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.8% agarose gel

The Southern blot was hybridised with the gel purified 250 bp *Hin*dIII/*Sac*II fragment from CL#034/5/KS+ (see restriction map, Figure 3.20) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Hin*dIII
Lane 2:	*Cla*I
Lane 3:	*Xba*I
Lane 4:	*Pst*I
Lane 5:	*Eco*RI

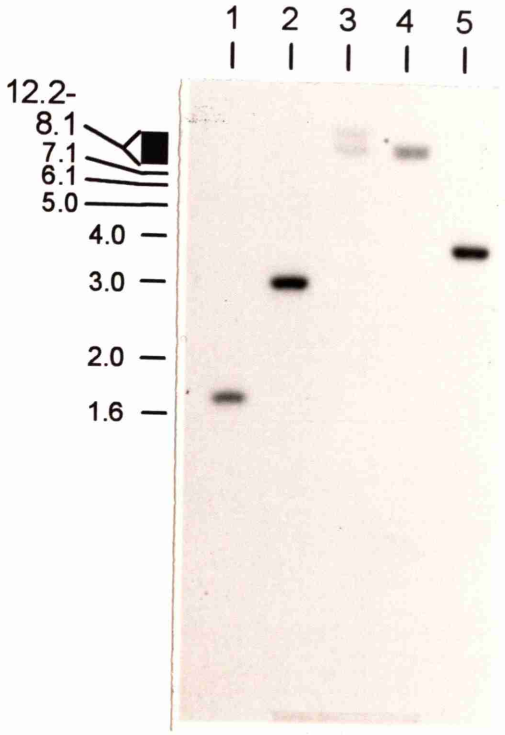

1
2
3
4
5
12.2-
8.1
7.1
6.1
5.0
4.0
3.0
2.0
1.6

Figure 3.30

A Southern blot of C25B8/10.7/KS- DNA digested with *Pst*I and *Sac*I and electrophoresed through a 1.0% agarose gel

The Southern blot was hybridised with pBluescript II KS- DNA and washed at high stringency prior to exposure with medical X-ray film.

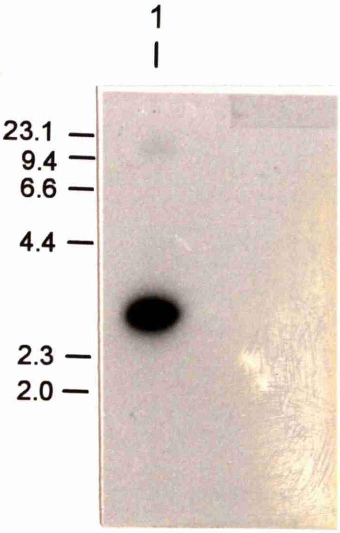

1
23.1
9.4
6.6
4.4
2.3
2.0

Figure 3.31

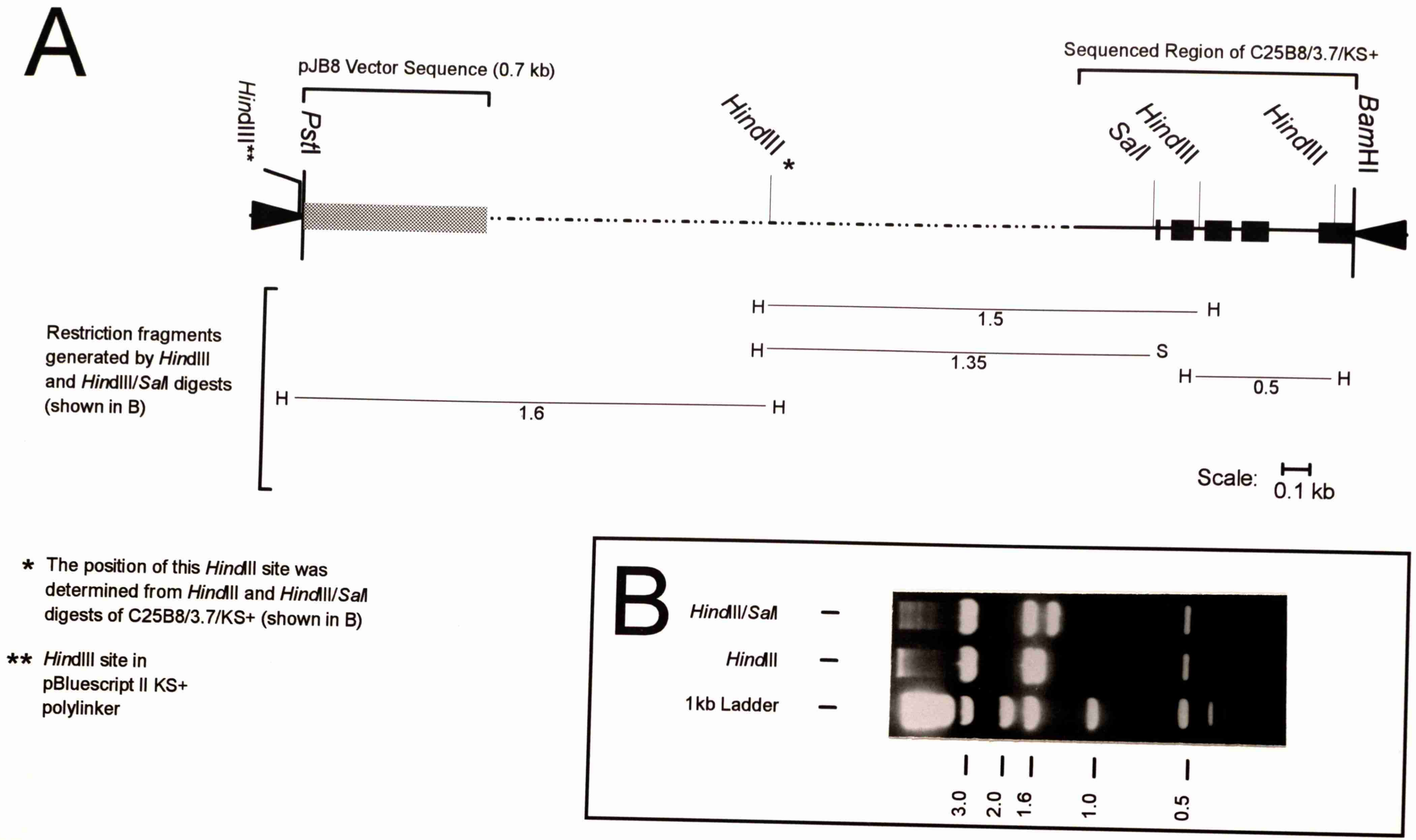
The restriction mapping data used to determine the position of the *Hind*III site upstream of the *cpr-6* gene in the C25B8/3.7/KS+ subclone

A. A summary of the positions of the *Hind*III sites in the C25B8/3.7/KS+ subclone

The *Hind*III sites were determined by nucleotide sequencing of the *cpr-6* gene and by restriction mapping using *Hind*III and *Hind*III/SalI digests.

B. The restriction map data used to determine the *Hind*III site upstream of the *cpr-6* gene

Note: Within the sequenced region of C25B8/3.7/KS+, the boxes indicate coding regions of the *cpr-6* gene and the lines between these boxes indicate introns.


A
pJB8 Vector Sequence (0.7 kb)
Sequenced Region of C25B8/3.7/KS+
HindIII**
PstI
HindIII *
SalI
HindIII
HindIII
BamHI
Restriction fragments generated by HindIII and HindIII/SalI digests (shown in B)
H 1.5 H
H 1.35 S
H 0.5 H
H 1.6 H
Scale: 0.1 kb
B
HindIII/SalI
HindIII
1kb Ladder
3.0
2.0
1.6
1.0
0.5

* The position of this *Hind*III site was determined from *Hind*III and *Hind*III/*Sal*I digests of C25B8/3.7/KS+ (shown in B)

** *Hind*III site in pBluescript II KS+ polylinker

Figure 3.32

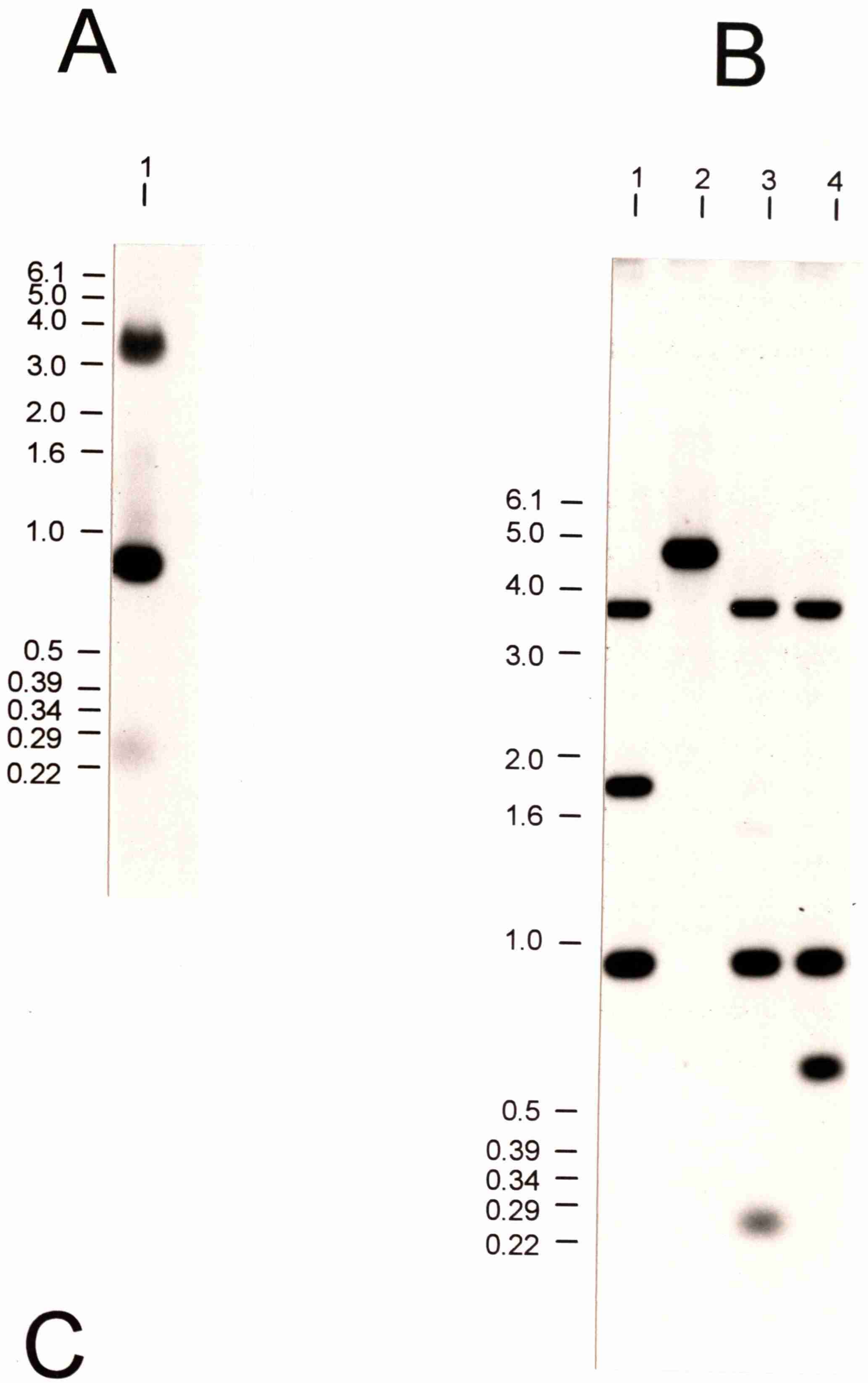
Diagnostic digests performed with *C.elegans* genomic DNA in order to determine whether the CL#034/5/KS+ and CL#034/1.7/KS+ subclones containing the *cpr-3* gene are contiguous with the *C.elegans* genome.

A. A Southern blot of *Eco*RV/*Sac*II/*Hin*dIII digested CL#034/5/KS+ DNA electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-3* cDNA (cm12b6) and washed at high stringency prior to exposure with medical X-ray film.

B. A Southern blot of *C.elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-3* cDNA (cm12b6) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Hin*dIII/*Xho*I
Lane 2:	*Xho*I/*Sac*II/*Eco*RV
Lane 3:	*Xho*I/*Sac*II/*Hin*dIII
Lane 4:	*Xho*I/*Eco*RV/*Hin*dIII

C. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in **B**



C

Lane Number	Digest	Expected Fragment Sizes
1	*Hind*III / *Xho*I	0.9 kb, 1.7 kb, 3.7 kb
2	*Xho*I / *Sac*II / *Eco*RV	4.85 kb
3	*Xho*I / *Sac*II / *Hind*III	0.25 kb, 0.9 kb, 3.7 kb
4	*Xho*I / *Eco*RV / *Hind*III	0.6 kb, 0.9 kb, 3.7 kb

Figure 3.33

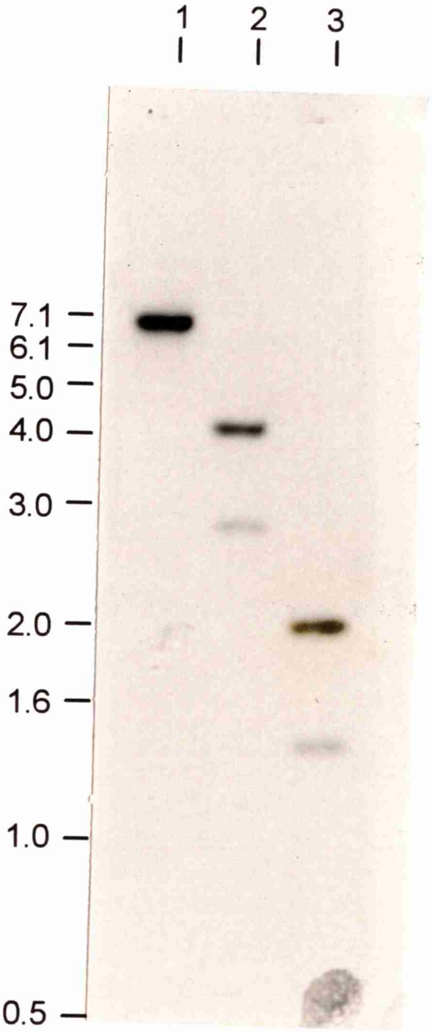
Diagnostic digests performed with *C.elegans* genomic DNA in order to determine whether the ZK1055/6.8/KS- and ZK1055/2.8/KS+ subclones containing the *cpr-4* gene are contiguous with the *C.elegans* genome.

A. A Southern blot of *C.elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-4* cDNA (cm14e3) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Hind*III
Lane 2:	*Hind*III/*Bam*HI
Lane 3:	*Pst*I/*Bam*HI

B. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in **A**

A



B

Lane Number	Digest	Expected Fragment Sizes
1	*Hin*dIII	6.8 kb
2	*HindIII* / *Bam*HI	2.9 kb. 4.1 kb
3	*PstI* / *BamHI*	1.4 kb, 2.0 kb

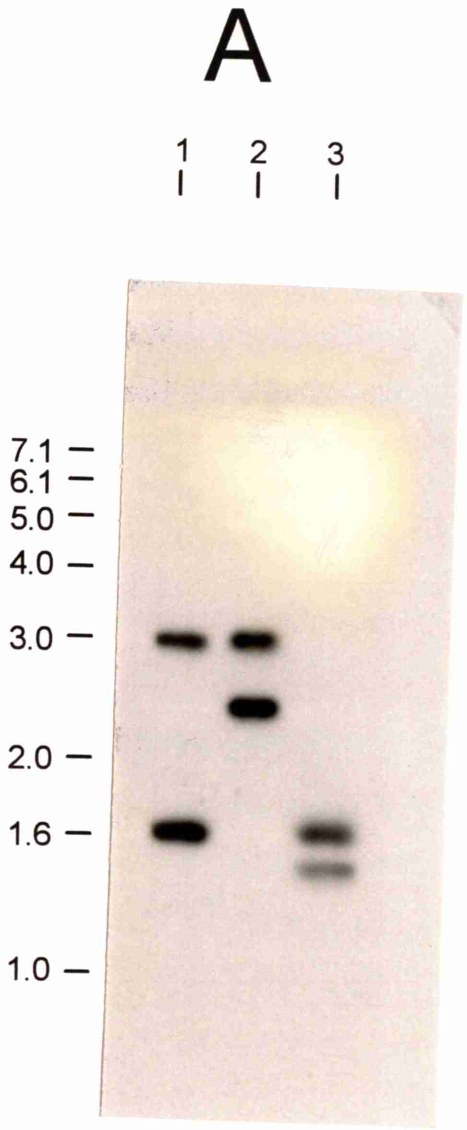
Figure 3.34

Diagnostic digests performed with *C.elegans* genomic DNA in order to determine whether the W02B2/3/KS+ and W02B2/1.6/KS+ subclones containing the *cpr-5* gene are contiguous with the *C.elegans* genome.

A. A Southern blot of *C.elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-5* cDNA (cm04d10) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Bam*HI/*Pst*I
Lane 2:	*Bam*HI
Lane 3:	*Bam*HI/*Pst*I/*Xba*I

B. Table summarising the expected sizes of the hybridising fragments generated from the diagnostic digests in **A**



B

Lane Number	Digest	Expected Fragment Sizes
1	*BamHI / PstI*	3.05 kb, 1.6 kb kb
2	*BamHI*	3.05 kb, and a fragment larger than 1.6 kb
3	*BamHI / PstI / XbaI*	1.6 kb, 1.4 kb

Figure 3.35

Diagnostic digests performed with *C.elegans* genomic DNA in order to determine the extent of the region of the C25B8/10.7/KS- subclone containing the *cpr-6* gene which is contiguous with the *C.elegans* genome.

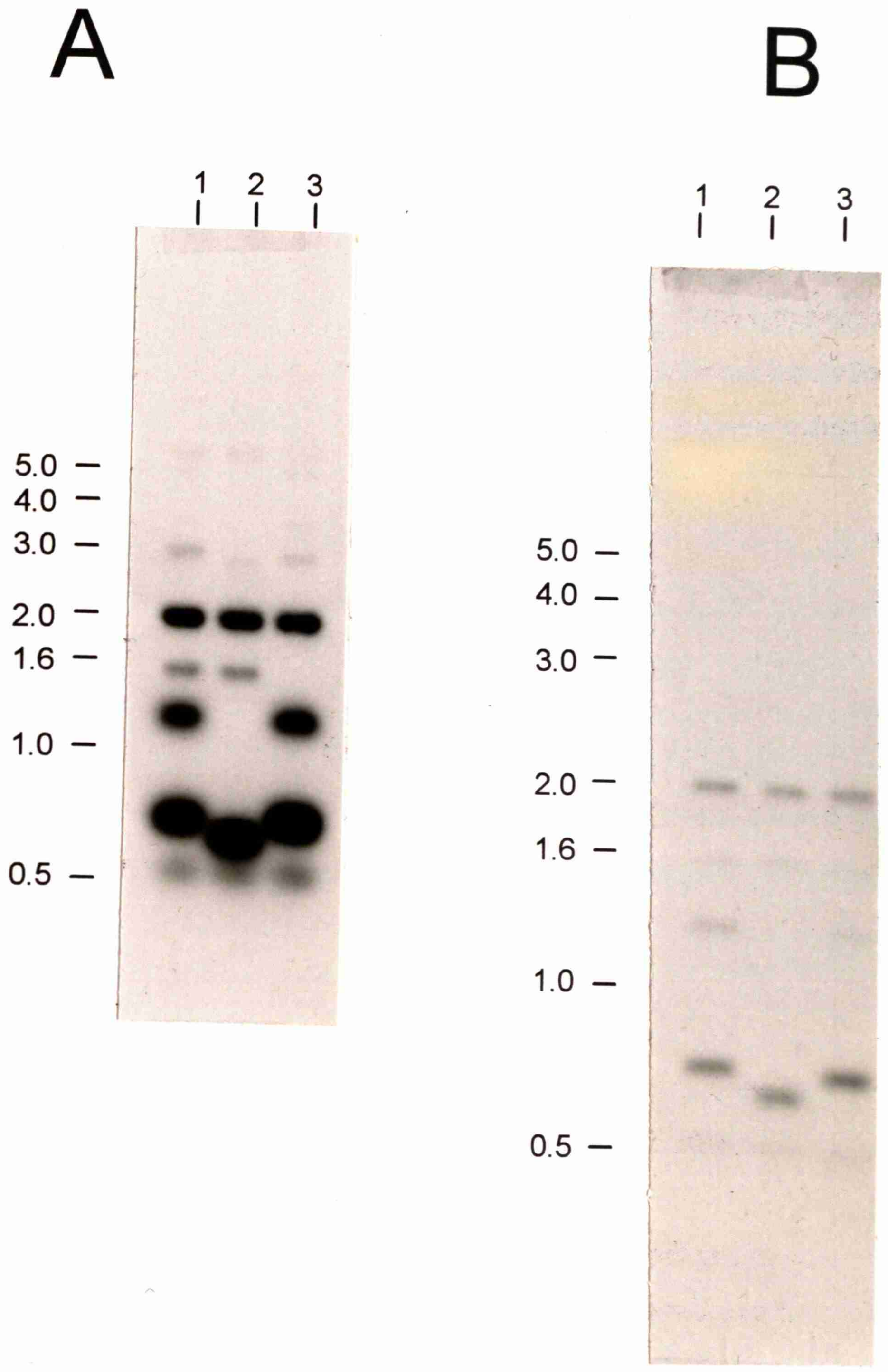
A. A Southern blot of C25B8/10.7/KS- DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-6* cDNA (cm01a5) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Hind*III
Lane 2:	*Hind*III/*Bam*HI
Lane 3:	*Hind*III/*Sal*I

B. A Southern blot of *C.elegans* Bristol N2 genomic DNA digested with a variety of restriction endonucleases and electrophoresed through a 0.7% agarose gel. The Southern blot was hybridised with the *cpr-6* cDNA (cm01a5) and washed at high stringency prior to exposure with medical X-ray film.

Lane 1:	*Hind*III
Lane 2:	*Hind*III/*Bam*HI
Lane 3:	*Hind*III/*Sal*I

C. Table summarising the expected sizes of the hybridising fragments obtained from the diagnostic digests in **A** and **B**



C

Lane Number	Digest	Expected Fragment Sizes
1	*Hind*III	0.5 kb, 0.7 kb, 1.5 kb, fragment of unknown size
2	*HindIII / Bam*HI	0.5 kb, 0.63 kb, 1.5 kb, fragment of unknown size
3	*Hind*III / *SalI*	0.5 kb, 0.7 kb, 0.15 kb, fragment of unknown size

Chapter 4

Chapter 4

Characterisation of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

4.1. Introduction

The isolation of four cathepsin B-like genes from *C.elegans* reported in Chapter 3, in addition to the previously isolated *cpr-1* gene (Ray and McKerrow, 1992), indicates that this free living nematode species possesses a cathepsin B-like multigene family comprising at least five members. Multigene families with the potential to encode cathepsin B-like enzymes have also been isolated from parasitic nematode and trematode species (Chapter 1, Section 1.9) but not from any other metazoan.

The presence of a cathepsin B-like multigene family in both parasitic and free living nematode species raises interesting questions as to the biological function, and degree of interaction, of the members of this gene family. Such a multigene family may have arisen as a result of a requirement for a large amount of cathepsin B-enzyme over a very short period of time. However, this seems unlikely since one might expect such a requirement to be fulfilled by a single cathepsin B-like gene under the control of a strong promoter. Therefore, a more likely explanation for the amplification of this multigene family in nematode species, may be the requirement for a range of cathepsin B-like enzymes with differing substrate specificities. The multicellular nature of nematode species may be an important factor in this respect, since enzymes with different substrate specificities may be required in different tissues, or in different compartments within tissues. Alternatively, such enzymes may be required at different stages during nematode development.

The subclones containing *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (Chapter 3) were isolated using four partially sequenced cDNA clones. The partial sequence data available for the four cDNA clones were originally generated by only a single sequencing reaction on one strand of each cDNA clone. Consequently, the amount (approximately 300 bp) and reliability of this sequence data were limited. Therefore, I decided to sequence the cDNA and genomic clones completely. This approach covered several objectives: to confirm that the subclones isolated using the four cDNA clones (Chapter 3) contained the genes corresponding to these cDNA clones; to determine the predicted primary protein structure of the four enzymes; to determine the genomic architecture of the four genes; to map the region of transcription initiation and to identify putative regulatory sequence elements in the 5′ flanks of the four genes.

4.2. Results

4.2.1. Confirming the identity of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The cDNA clones were sequenced completely on both strands. The regions of each gene within the appropriate subclone that were also contained within the cDNA clones were sequenced on one strand. A comparison of the cDNA and coding genomic sequences revealed that the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 shared complete base identity to their genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6,* present in the genomic clones CL#034, ZK1055, W02B2 and C25B8, respectively. These data formally indicated that the correct genes had been isolated using these cDNA clones as probes. Therefore, the cDNA clones cm12b6, cm14e3, cm04d10 and cm01a5 are subsequently referred to by their gene names, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. A more detailed discussion of the data revealed by this comparison will be given later.

4.2.2. Sequencing the cDNA clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

Each of the four cDNA clones were sequenced on both strands. The results of this analysis indicated that all four clones possess intact 3′ ends with poly(A) tails of at least eleven residues. The cDNA clones of *cpr-3* and *cpr-4* both possessed canonical cleavage/poly(A) signals (AAUAAA) 14 bp and 15 bp upstream of the poly(A) tail. The two other cDNAs of *cpr-5* and *cpr-6* possess variations to this signal of GAUAAA and AAUGAA respectively, 15 bp and 13 bp upstream of the poly(A) tail, which probably serve as the cleavage/poly(A) signal. This conclusion is drawn from a survey of the 3′ ends of 1303 *C.elegans* cDNA clones (T.Blumenthal, pers. comm.). This survey revealed that a significant number of the cDNA clones analysed have variations of the canonical cleavage/poly(A) signal upstream of the poly(A) tail. Two of the more common variations found are GAUAAA (4.8%) and AAGUAA (8.3%). The survey also revealed that putative cleavage/poly(A) signals are most frequently found 11-15 bp upstream of the poly(A) tail. These observations lead T.Blumenthal (pers. comm.) to suggest that any of the sequence variations found with a frequency of over 1% in the correct locale are likely to be acting as cleavage/poly(A) signals.

The predicted amino acid sequence (discussed in detail in Section 4.2.7) indicated that the cDNA clones of *cpr-3*, *cpr-4* and *cpr-6* probably contained the ATG translation initiation codons, however the 5′ untranslated regions (UTRs) of these three clones were short (21 to 39bp), suggesting that these clones may be truncated at the 5′ end. The cDNA clone of *cpr-5* is definitely truncated at the 5′ end, with no putative ATG initiation codon present.

4.2.3. Obtaining the full-length mature transcripts of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

I decided to use the 5′ Rapid Amplification of cDNA Ends (5′ RACE) method to obtain the 5′ ends of the mature transcripts for each of the four genes. A summary of this protocol is shown in Figure 4.1, and the nucleotide sequences of all the primers used are given in Chapter 2, Section 2.11.3.2. Two pools of first strand cDNA were generated from *C.elegans* Bristol N2 mixed stage total RNA using pooled gene specific primers for each of the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* and Superscript II Reverse Transcriptase (GibcoBRL). The two pools of first strand cDNA were tailed using dCTP Terminal Deoxynucleotidyl Transferase (TdT) and amplified by 36 cycles of PCR using pooled nested gene specific PCR primers for each of the four genes and the GibcoBRL anchor primer, according to the GibcoBRL 5′ RACE protocol. Each of the two sets of amplification products generated were re-amplified in four independent PCR reactions, using the anchor primer, one nested gene specific PCR primer and 36 cycles of PCR. The results obtained using 5′ RACE for each of the four genes are show in Figure 4.2. The results indicated that the 5′ RACE reactions from the two pools of cDNA reproducibly generate the same size of predominant products for each gene which migrate as discrete bands when electrophoresed through a 1.5% agarose gel. The 5′ RACE experiments generated a single predominant size of product for *cpr-3*, *cpr-4* and *cpr-5*, suggesting that the 5′ ends of the mature transcripts of these genes are predominantly of a single species. For *cpr-6*, two predominant sizes of 5′ RACE product are generated, migrating approximately 60 bp apart, initially suggesting that there may be two predominant species of mature transcript derived from this gene. The predominant products were gel purified and cloned into pBluescript II KS+ by virtue of restriction enzyme sites present in the primers used for the final amplification reaction. A total of at least five 5′ RACE clones, generated from the two cDNA populations, were sequenced on both strands for each gene. The nucleotide sequence of each 5′ RACE clone was determined for the region outwith the 5′ end of the appropriate cDNA clone, in addition to a 50 bp overlap to confirm identity with the cDNA clone sequence. The cloning sites at the 5′ and 3′ ends of each 5′ RACE clone were then identified by scanning the remaining sequence of each clone. This allowed the size of each 5′ RACE clone to be determined, and these sizes were compared to the expected sizes obtained from initial electrophoresis of the 5′ RACE products (Figure 4.2). This comparison (Table 4.1) reveals that the sizes of the cloned 5′ RACE products correlate well with the expected sizes determined from agarose gel electrophoresis of the 5′ RACE products prior to cloning.

For *cpr-6*, two similarly abundant 5′ RACE products were obtained, migrating approximately 60 bp apart. These products were gel purified together and cloned into pBluescript II KS+. Six of the cloned products were selected at random and sequenced. These results indicated that all six clones were of a similar size to each other, and corresponded to the smaller of the two products generated by 5′ RACE (Table 4.1). This was a surprising result since the two 5′ RACE products were of a similar abundance, and therefore one would expect both products to be represented in the 5′ RACE clones selected for sequence analysis. These data suggested that the 5′ end of the mature transcript of *cpr-6* was also predominantly of a single species, and therefore the larger product generated after the second round of PCR was probably an artefact. The second round of PCR was repeated for *cpr-6* using one of the two pools of amplified cDNA previously obtained from the first round PCR. The results indicated that only a single predominant 5′ RACE product is generated (Figure 4.3), which is a similar size to the smaller of the two products previously observed for *cpr-6* (Figure 4.2). These data support the assumption that the larger 5′ RACE product previously observed for *cpr-6* was indeed an artefact.

The results obtained from DNA sequencing the cloned 5′ RACE products of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are shown in Figures 4.4, 4.5, 4.6 and 4.7, respectively. Some of the products showed base substitutions at various points in the sequence and these have been indicated. These substitutions were assumed to have occurred during the PCR reaction. The *Taq* DNA polymerase used in the PCR reaction does not have any proof reading ability, and therefore base mismatches may occur during the amplification step. Furthermore, the frequency of such errors increases with the number of PCR cycles used. In this instance, the use of a large number of PCR cycles (72 in total) was therefore likely to generate several errors. In order to overcome this a consensus sequence was generated from all the sequence data obtained from the cloned 5′ RACE products, for each gene. A comparison of the consensus sequence obtained from the 5′ RACE products for each of the genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to the appropriate cDNA revealed complete base identity in the region of overlap, with the exception of the 5′ most base of each of the cDNA clones. This base was immediately downstream of the *Sac*I site used to clone these cDNAs during initial preparation of the library (Palazzolo *et al.,* 1990) and was therefore probably an artefact. The library was originally prepared by size fractionation of poly(A)$^+$ RNA, generating first strand cDNA with an oligo(dT)-*Apa*I primer-adapter and 5-methyl-dCTP instead of dCTP. Second strand cDNA was then generated using unmethylated dCTP, to generate hemimethylated cDNA. A *Sac*I linker was then ligated to the 5′ end of the cDNA and the cDNAs digested with *Apa*I and *Sac*I (any endogenous sites being protected by the methylation) and ligated into the lambda SHLX2 phage vector. All the cDNA clones possess a guanine residue immediately downstream of the *Sac*I site, and it is this residue in addition to the *Sac*I site, that mismatches with the sequence data obtained from the 5′ RACE clones. This residue was therefore probably added to each of the cDNA clones during the generation of the cDNA library, as part of the *Sac*I linker.

The sequence data obtained from the cloned 5′ RACE products of *cpr-3* (Figure 4.4), *cpr-4* (Figure 4.5), *cpr-5* (Figure 4.6), and *cpr-6* (Figure 4.7) all generated additional sequences outwith the 5′ end of the corresponding cDNA clone. For *cpr-5* (Figure 4.6), the 5′ RACE products generated a single ATG codon which was in frame with the predicted amino acid sequence of the corresponding cDNA clone. This ATG codon probably represents the translation initiation codon, since eukaryotic translation is thought to occur at the first ATG codon of a given transcript (Kozak, 1983). For *cpr-4* (Figure 4.5), and *cpr-6* (Figure 4.7), translation is most likely to be initiated from the ATG codons previously identified in the cDNA clones (Section 4.2.2), since the sequenced 5′ RACE clones from these two genes did not reveal any additional ATG codons in the 5′ UTRs of their mature transcripts. For *cpr-3* (Figure 4.4), the ATG codon originally identified by sequencing the cDNA clone of this gene (4.2.2) is also the putative translation initiation codon. However, one of the eight sequenced 5′ RACE clones of *cpr-3* possesses an additional ATG codon upstream of, and in-frame with, this putative translation initiation codon. This upstream ATG codon is unlikely to be the translation initiator since an in-frame translation termination codon separates it from the putative translation initiator. Thus, translation initiation from this ATG codon would only generate a 22 amino acid peptide. It is possible that this peptide may have a regulatory role such as an enzyme inhibitor, or by preventing synthesis of the enzyme from this transcript. However, the physiological significance of this additional ATG codon must be questioned since only one of the eight sequenced 5′ RACE clones of *cpr-3* possesses it, suggesting that this ATG codon is not present in the majority of mature transcripts generated from the *cpr-3* gene.

The predicted amino acid sequences, generated from the predicted translation initiation codons for each of the four genes, all generate very similar primary protein structures (Figure 4.8, discussed later). In particular, all four possess highly hydrophobic putative pre-enzyme domains which have been identified in all other cathepsin B-like genes sequenced to date. The generation of four structurally similar primary protein sequences from the mature transcripts of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, using the chosen ATG codons, strongly supports their role as translation initiators.

The cloned 5′ RACE products for all four genes, *cpr-3*, *cpr-4, cpr-5* and *cpr-6*, all indicated short 5′ UTRs. A comparison of the consensus sequence of the 5′ UTRs, up to and including the ATG translation initiation codon, to the 5′ flank sequence obtained from the genomic clones for each of the four genes revealed complete base identity along their entire length (Figure 4.9). This indicates that none of the predominant transcripts from *cpr-3*, *cpr-4, cpr-5* and *cpr-6* are trans-spliced to either of the two 22 nucleotide splice leaders of *C.elegans*, SL1 and SL2 (Appendix A). This is unusual since recent reports suggest that approximately 70% of *C.elegans* genes are trans-spliced to one of these two splice leaders (Zorio *et al.,* 1994). However, these data are in agreement with data obtained from the previously isolated cathepsin B-like gene of *C.elegans*, *cpr-1*, which also suggest that this gene is not trans-spliced (Ray and McKerrow, 1992). These data suggest that the members comprising the cathepsin B-like multigene family of *C.elegans* all lack this aspect of transcript maturation. In *C.elegans*, some genes have been shown to be transcribed polycistronically in clusters resembling bacterial operons (Spieth *et al.,* 1993). Mature, monocistronic messages are subsequently generated from the polycistronic message by trans-splicing events. A recent study of potential operons revealed by the *C.elegans* genome sequencing project suggested that approximately 26% of genes may be organised into operons (Zorio *et al.,* 1994). The authors demonstrated that, of the operons studied, the first gene was either trans-spliced to SL1 or not trans-spliced at all while the downstream genes were exclusively trans-spliced to SL2, or a mixture of SL1 and SL2. These data suggest that if any of the *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* or *cpr-6* genes are transcribed from *C.elegans* operons, they will represent the first gene in their operon. The two genes, *cpr-1* and *cpr-3*, are closely linked on chromosome V (Chapter 3, Section 3.2.13) suggesting that, of all the cathepsin B-like genes identified to date, these two are the most likely to be co-transcribed from the same operon. However, the data suggest that neither of these two genes are trans-spliced and are therefore unlikely to represent the downstream gene of a *C.elegans* operon. This suggests that these two genes cannot be transcribed from the same operon, though they could each represent the first gene of two different operons.

4.2.4. Analysis of the full-length mature transcripts of *cpr-3*, *cpr-4*, *cpr-5 cpr-6*

The full length mature transcripts were obtained from the combined sequence data of the cDNA and 5′ RACE clones for each gene. The sizes of the transcripts were calculated from the largest 5′ UTR generated by more than one 5′ RACE clone (for *cpr-3*, *cpr-4* and *cpr-6*). For *cpr-5*, the largest 5′ UTR was used because none of the sequenced 5′ RACE clones terminated at the same position for this gene. The sizes were found to be, 1,329 bases, 1,120 bases, 1,152 bases and 1,260 bases for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively. The open reading frames of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encode proteins of 370, 335, 344 and 379 amino acids respectively. Subsequent analysis of these amino acid sequences indicated that these four genes encode cathepsin B-like cysteine protease enzymes that are similar to, but distinct from, each other and the 329 amino acid cathepsin B-like gene product of the previously isolated *cpr-1* gene, a more detailed comparison of these sequences will be given later. The mature transcripts of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* all shared similar features. As discussed in the previous section, none of the transcripts are trans-spliced. Furthermore, the 5′ untranslated sequences of these five genes are all relatively small. The 5′ RACE experiments performed for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* generated several 5′ UTRs of slightly different length for each gene (Figures 4.4, 4.5, 4.6 and 4.7). The size ranges for the 5′ UTRs of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* obtained from the 5′ RACE experiments are 39-92 bases, 33-39 bases, 14-29 bases and 29-46 bases, respectively. Only the mature transcript of *cpr-3* possesses a 5′ UTR substantially larger than is found for the other three genes and *cpr-1*. It should be noted that the wide range of 5′ UTR sizes found for *cpr-3* are slightly misleading since the extremities of these sizes (92 bases and 39 bases) were each generated by single 5′ RACE clones. The majority of 5′ RACE clones for *cpr-3* generated 5′ UTRs of 60 bases (3 clones) and 55 bases (3 clones) which were only slightly larger than the 5′ UTRs found for the other *C.elegans* cathepsin B-like genes. The 5′ UTRs of the mature transcripts of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are a similar size to the 28 bp 5′ UTR of *cpr-1*, determined using primer extension analysis (Ray and McKerrow, 1992). These data suggest that short 5′ UTRs may be a common feature of the mature transcripts of all *C.elegans* cathepsin B-like genes.

The FOLDRNA program of the GCG package was used to predict RNA secondary structures for the 5′ and 3′ UTRs of *cpr-3*, *cpr-4*, *cpr-5*, *cpr-6* and *cpr-1* in order to determine whether there were any stable stem-loop structures in the 5′ UTRs that could potentially perform regulatory roles for these genes. Only the 3′ UTR of *cpr-3* generated a stem-loop structure that showed significant stability (-20.1 kilocalories at 37°C, Figure 4.10). This is not surprising since the 3′ UTR of *cpr-3* (156 bases) is notably larger than the 3′ UTRs of *cpr-4* (76 bases), *cpr-5* (88 bases), *cpr-6* (75 bases) and *cpr-1* (62 bases). The stable stem-loop structure generated by the 3′ UTR of the *cpr-3* transcript may represent a mechanism by which *cpr-3* expression is post-transcriptionally regulated (discussed in Section 4.3.4.1).

4.2.5. The gene architectures of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The genomic copies of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were sequenced using custom made primers (Chapter 2, Section 2.11.3.2) and template DNA from the following subclones: CL#034/1.7/KS+, CL#034/3/KS+ and CL#034/5/KS+ for *cpr-3*; ZK1055/2.8/KS+ and ZK1055/6.8/KS- for *cpr-4*; W02B2/1.6/KS+ and W02B2/3/KS+ for *cpr-5*; C25B8/3/KS+, C25B8/3.7/KS+ and C25B8/10.7/KS- for *cpr-6*. A summary of the position of these subclones with respect to one another and with respect to the appropriate gene is shown in Chapter 3, Figures 3.25, 3.26, 3.27 and 3.28. It should be noted that W02B2/1.6/KS+ and W02B2/3/KS+ are adjacent, not overlapping subclones. These clones were shown to be adjacent by comparison of the nucleotide sequence of the *cpr-5* gene, which spans these two clones, to the nucleotide sequence of the cDNA clone of the *cpr-5* gene (cm04d10).

The coding sequences of the genomic clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were determined by sequencing one strand and confirming identity with the corresponding cDNA clone. Those regions of the genomic sequence not contained within the cDNA clones (introns and 5′ and 3′ flanks), were sequenced on both strands. A comparison of the nucleotide sequence of the mature transcripts (generated from the combined cDNA clone and 5′ RACE product sequence data) of each of the four genes to the nucleotide sequence of their genomic clones, reveals complete base identity interrupted by introns. The annotated DNA sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* generated by these comparisons are shown in Figures 4.11, 4.12, 4.13 and 4.14, respectively. All the introns revealed within *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* possess the conserved splice donor (GT) and splice acceptor (AG) sites expected for *C.elegans* introns (Emmons, 1988). However; the size, number and position of these introns vary considerably between the genes giving rise to quite different gene architectures, which are shown schematically in Figure 4.15. Intron size ranges from 43 bp (*cpr-6*) to 406 bp (*cpr-3*) while number varies from one (*cpr-4*) to seven (*cpr-6*). The range in size and number of introns for the genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6,* are well within the limits expected for *C.elegans*. Early analysis of 54 different *C.elegans* introns revealed that three-quarters were between 45 and 59 bases long (Emmons, 1988). Experiments performed with the rabbit ß-globin gene suggest introns shorter than 80 bp may not be effectively spliced in higher eukaryotes (Wieringa *et al.,* 1984). This suggests that the mechanism employed for splicing *C.elegans* introns is different from that for splicing introns from higher eukaryotes.

The position of the introns were also compared at the amino acid sequence level (Figure 4.8). In most cases intron position at the amino acid level is not conserved, however there are two exceptions. The *cpr-4* and *cpr-6* genes share a common intron site immediately after the ATG start codon (position 1) with the previously characterised gene *cpr-1*. The *cpr-3* and *cpr-5* genes share a common intron site with *cpr-1* at position 261 (Figure 4.8). These intron positions are not only conserved with respect to the amino acid sequence but also with respect to the exact nucleotides at which intron / exon boundaries occur within the codons. The identification of two conserved sites of intron position with respect to the amino acid and nucleotide sequences of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6,* suggest that these genes share a common ancestor. However, the different gene architectures, in terms of the position, number and size of the introns, suggest that these genes are highly diverged.

The COMPARE and DOTPLOT programs of the GCG package were used to identify homologous sequences from different regions of the same gene which might represent repeated sequences, by comparing *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to themselves. These programs were also used to identify regions of homology between the genes by comparing the nucleotide sequences of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to each other. In the latter case, the COMPARE and DOTPLOT programs were necessary because the DNA sequences of each of the five genes are sufficiently diverged such that the GAP program of the GCG package (which aligns sequences by introducing gaps) was unable to generate any alignments between the different genes.

The COMPARE program works by scanning the sequence of one gene for homology to a window of sequence taken from the same, or another gene. The window of sequence is then moved along that gene in a stepwise manner and the process repeated. Regions of homology are then scored on a graph as a dot, using the DOTPLOT program. Successive regions of homology therefore give rise to successive dots which generate lines whose length are proportional to the length of the stretch of homology.

DOTPLOT comparisons of each of the genes *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to themselves did not reveal any significant homologies between different regions of the same gene, except for *cpr-5*. For this gene, extensive homology was found in the first 80 bp of the 5′ flank of the sequenced gene. A study of this region revealed several repetitions of two hexamer sequences, TTATGC and TTAGGC (Figure 4.16). Seven repeats of TTATGC and four of TTAGGC were identified. However, further repeats containing one base pair mismatches to these sequences were also identified interspersed amongst these perfect repeats. The TTAGGC repeat identified is thought to be the telomeric hexamer of *C.elegans*. Recently the complete set of all 12 telomeres of *C.elegans* has been analysed and each telomere shown to comprise 4-6 kb of this repeated hexamer (F.Muller, pers. comm.). Interstitial repeats of TTAGGC have also been found in *C.elegans*, and these tend to be clustered in the terminal 30% of chromosomes (Cangiano and La Volpe, 1993). In such cases, degenerated copies of the telomeric hexamer have been observed, interspersed among perfect copies of this repeat sequence (Cangiano and La Volpe, 1993). The type and pattern of repeat sequence found in the 5′ flank of *cpr-5* suggests that this gene is closely associated with a region of interstitial telomeric repeats. This is not surprising since *cpr-5* has been mapped to the left end of chromosome V (Chapter 3, Section 3.2.13), where many interstitial copies of the telomeric repeat appear to be clustered. It should be noted that this region of the 5′ flank of *cpr-5* was not sequenced on both strands because it was not possible to generate a specific primer, required for sequencing the second strand within the region of repetitive sequence. Therefore, it was not possible to determine whether the degenerated copies of the two types of hexamer repeat found were genuine or a result of sequencing inaccuracies. Though the two types of hexamer repeat only differ themselves by one nucleotide residue, this variation is unlikely to be a result of sequencing inaccuracies because several copies of each type of repeat were observed.

The COMPARE and DOTPLOT programs identified several regions of homology between *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. A summary of the data obtained from these dotplots (the position and length of the homologies) is given in Table 4.2. When these positions are compared to the annotated sequence data of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (Figures 4.11, 4.12, 4.13 and 4.14, respectively) it becomes apparent that these homologies are within the regions of the genes coding for the catalytic residues (cysteine, histidine and asparagine), and the amino acid residues surrounding these sites. These data suggested that these genes are sufficiently diverged such that no significant DNA homology is present between introns, or exons outwith the regions which code for the catalytic domains of their gene products. The stretches of homology between genes, identified by the dotplots and summarised in Table 4.2, were compared to determine the degree of homology to each other using the GAP program of the GCG package. Though the GAP program was unable to generate alignments between the complete sequences of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, it was able to generate alignments between the regions of homology identified from the dotplots. Presumably, this was a result of the fact that much shorter DNA sequences were being compared and that these sequences had already been selected on the basis of increased homology to one another using the COMPARE and DOTPLOT programs. The results of the GAP program analysis (summarised in Table 4.3) demonstrate that whilst all the genes share regions of significant homology to each other (80.8% - 100%), the length of these homologies are very small, the maximum being 29 bp of continuous homology between *cpr-1* and *cpr-4*. The lack of any continuous homology over 29 bp between any two genes suggested that the genes *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are highly diverged from each other at the nucleotide level. Indeed, the degree of divergence is sufficiently high to suggest that detection of one gene, using another as a hybridisation probe, would be difficult, or impossible.

Restriction endonuclease fingerprinting previously demonstrated that the clones containing *cpr-1* and *cpr-3* are very closely linked, with the two clones mapping on top of one another (Chapter 3, Section 3.2.13). Therefore, the COMPARE and DOTPLOT programs were used to compare the 5′ and 3′ flanks of *cpr-3* to those of *cpr-1* (in both orientations). This comparison did not identify any regions of homology between the 5′ and 3′ flanks of *cpr-1* and *cpr-3*, indicating that neither of these two genes possess sequences of the other in the sequenced regions of their 5′ and 3′ flanks.

4.2.6. Putative regulatory elements in the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

5′ RACE was used initially to obtain the predominant full size transcripts of the genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. However, the discovery that these genes were not trans-spliced to either SL1 or SL2, allowed the 5′ RACE results to be used to map the general region of transcription initiation for each of the genes. The 5′ UTRs defined by the cloned 5′ RACE products, up to and including the ATG translation initiation codon, were compared to approximately 300 bp of 5′ flank sequence of the appropriate gene. This comparison demonstrated that the 5′ ends of the RACE clones from each gene clustered to a small region of the 5′ flank of the appropriate gene (Figure 4.17). This tight clustering suggests that transcription initiation occurs within one region of the 5′ flank of each of the four genes. The only slight exception to this is in the case of *cpr-3*, where the 5′ ends of two 5′ RACE products were separated by 52 bases. In this example, however, the 5′ ends of the six other 5′ RACE clones for *cpr-3* mapped to two positions five bases apart. These data suggested that transcription initiation occurs predominantly from a single region, but that spurious transcription initiation may also occur outwith this area.

A survey of the 5′ flank sequences of the four genes upstream of the regions to which the 5′ ends of the 5′ RACE clones mapped, revealed clustering of several putative regulatory elements. In two cases, *cpr-4* and *cpr-6*, putative TATA box (TATAWAW) promoter elements were identified, 25 and 26 bases respectively, upstream of the point to which most of their 5′ RACE clones mapped (Figure 4.17). Since TATA boxes are generally found 25 bp upstream of the region of transcription initiation, these results suggest that the TATA sequences identified may represent functional promoter elements. For *cpr-3* and *cpr-5*, no such potential TATA box promoter sequences were identified indicating that a TATA-less promoter must be involved in transcription of these two genes. All four genes possess sequences with homology to the motif recognised by the GATA family of transcription factors. The 5′ flank sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* contain four, five, two and two copies, respectively, of the classic (WGATAR, Figure 4.17) sequence recognised by the GATA family of transcription factors, plus additional related variants which might also bind this family of transcription factors (Ko and Engel, 1993; Merika and Orkin, 1993).

4.2.7. *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encode proteins with homology to cathepsin B

The predicted amino acid sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6,* and that of *cpr-1*, most closely resemble the cathepsin B cysteine proteases, as typified by human cathepsin B (Chan *et al.,* 1986). The predicted amino acid sequences of these six genes are compared in Figure 4.8. Each of the six genes possesses a short (16 - 19 residue) N-terminal hydrophobic domain. This hydrophobic domain is the candidate signal prepeptide region which may be responsible for directing the preproenzymes to the lumen of the endoplasmic reticulum during translation by the ribosomes (Docherty and Steiner, 1982). This domain is followed by a diverged region of 62 - 87 amino acids which is thought to represent the proregion of the enzymes. Such divergence in the proregion has also been noted between human cathepsin B and its homologues in the mouse and the rat (Chan *et al.,* 1986) and between the CP-1 and AC-2 cathepsin B-like enzymes of *H.contortus* and *O.ostertagi* (Pratt *et al.,* 1992b). The function of the proregion is unclear, however removal of this region is thought to be necessary for activity of the mature enzyme. Extracellular, stable, high molecular mass forms of recombinant human cathepsin B have been isolated from yeast. These forms appear to represent non-covalent complexes between the mature cathepsin B enzyme and its propeptide (Mach *et al.,* 1994b). The authors also reported that this complex does not exhibit a significant ability to cleave protein substrates. Furthermore, experiments using a synthetic rat cathepsin B propeptide have demonstrated that this propeptide is a potent reversible inhibitor of the rat cathepsin B enzyme (Fox *et al.,* 1992). The proregion of all six of the predicted protein products compared lack the interspersed ERFNIN motif. This motif has been identified in the proregion of a number of cysteine proteases but is absent from the proregion of all cathepsin B-like enzymes analysed (Karrer *et al.,* 1993). The predicted region of cleavage of the propeptide from the mature enzyme of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* is indicated in Figure 4.8. This position is based on the predicted site of cleavage for *cpr-1* (Ray and McKerrow, 1992) and the known site of cleavage for human cathepsin B (Chan *et al.,* 1986) which was determined by comparison of the predicted amino acid sequence obtained from a human cathepsin B cDNA and amino acid sequencing of mature human cathepsin B. The predicted site of cleavage of the propeptide from the mature enzyme of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* is very near to the site of cleavage for *AC-4* of *H.contortus*, determined by N-terminal protein sequencing (Pratt *et al.*, 1992a).

The predicted mature forms of the enzymes show much higher levels of homology than the prepropeptides, as can be seen from the number of amino acid residues conserved between the prepro- regions and the predicted mature regions of the six enzymes (Figure 4.8). The percentage similarities and identities between the prepro and predicted mature forms of the enzyme products of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5*, *cpr-6* and human cathepsin B are shown in Table 4.4. These data also demonstrate that the predicted mature forms of the enzymes are more homologous to one another (51.0% - 77.3% identity) than the preproenzyme forms (43.1% - 67.6% identity). However, the data summarised in Table 4.4 also demonstrate that the putative enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are not highly homologous, since (with the exception of *cpr-4* and *cpr-5*) these putative enzymes do not show significantly more homology to one another than to human cathepsin B.

Despite the diverged nature of the cathepsin B-like enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, the predicted mature forms of these enzymes possess the characteristics expected of cathepsin B enzymes. The three definitive catalytic residues, Cysteine (C), Histidine (H) and Asparagine (N), are highly conserved, as are the amino acid residues surrounding these sites. In addition, the position of these three catalytic residues, with the cysteine residue situated towards the N-terminus and the histidine and asparagine residues situated towards the C-terminus, are highly conserved between the six enzymes. Fourteen of the cysteine residues in the predicted mature enzymes can be aligned with few gaps. Since cysteine residues are required for the production of disulphide bonds, necessary for generating the tertiary structure of the proteins, the ability to align such a high number of these residues with a minimal number of gaps suggests these six gene products all have similar tertiary structures. These data indicate that the four genes isolated, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, encode protein products which, though not highly homologous to one another, display the hallmarks of the cathepsin B class of proteolytic enzymes, and therefore support the naming of these genes as cathepsin B-like.

4.2.8. Phylogenetic analysis

Phylogenetic analysis requires that the sequences to be analysed are aligned accurately. The Clustal program, version V (Higgins *et al.,* 1992) included in the PHYLIP package version 3.5c (Felsenstein, 1993). was used to generate all alignments for phylogenetic analysis. These alignments were checked by eye to ensure that anomalous gaps had not been introduced into the protein sequences to be analysed. Once the sequences had been aligned, two distinct methods, parsimony and distance-matrix, were used to estimate the evolutionary trees from these alignments, using the PROTPARS and NEIGHBOR programs, respectively, of the PHYLIP package version 3.5c (Felsenstein, 1993).

The PROTPARS method, like all parsimony approaches, can be used to infer phylogenies directly from the character data obtained from the alignment. In essence this means scoring each character (amino acid residue) for similarity to characters in homologous positions from the other sequences. As is suggested by their names, parsimony algorithms assume that the shortest evolutionary trees are the nearest approximation to the 'true' trees. Thus these programs select the trees that minimise the total tree length (i.e. minimise the number of evolutionary steps, or transformations from one character state to another, required to explain a given data set). The PROTPARS program uses several constraints when making this comparison. First, it insists that any changes of amino acid be consistent with the genetic code. However changes between two amino acids via a third are allowed and counted as two changes. Second, the algorithm does not score synonymous changes, nucleotide changes which do not alter the amino acid sequence. This is because DNA sequence comparisons suggest that synonymous changes are considerably faster and easier to make than ones that change the amino acid sequence. The algorithm therefore assumes that such changes need not be counted.

The best way to obtain an optimal tree, defined as the shortest tree when using parsimony analysis, is to determine all the possible trees that can be generated from a given dataset and subsequently select the shortest tree. Such exhaustive methods become impractical with large datasets, since they require substantial amounts of computer processor time. Heuristic approaches are therefore used which sacrifice the guarantee of optimality in favour of reduced computing time. The PROTPARS program is one example of many heuristic approaches which can be used to recreate a phylogenetic tree. This program recreates phylogenetic relationships by stepwise addition. Essentially this involves selecting three taxa (or sequences) to generate an initial tree. In the next step, one of the as yet unplaced taxa is added to the tree. In this process, each of the three trees that would be generated by adding this next taxon (or sequence) along one of the three branches, are evaluated and the optimal tree selected for the next round of addition. The process is then repeated for the next unplaced taxon until all the taxa have been placed on the tree. Thus, at each round of addition of a new taxon or sequence, the PROTPARS program selects the optimal tree (the shortest tree) prior to adding a new taxon. With such a method, locally optimal solutions will be obtained but there is no guarantee that the resulting tree will represent the globally optimal solution. However, the success of such a method in obtaining a globally optimal solution can be tested statistically. The bootstrap resampling technique (Efron and Gong, 1983) is one example of such a statistical test and was used to analyse some of the evolutionary trees recreated by the PROTPARS program. This technique takes random samples from the initial dataset (the alignment) and generates a new tree in each case. This process is repeated several times, and the number of times each grouping in the original tree occurs in the bootstrap trees is counted. These figures can then be used as a guide to the reliability of the clustering observed in the tree.

A distance-matrix approach was also used to infer phylogenetic relationships from the aligned sequences. When using such an approach the protein sequence alignments must first be transformed into pairwise-distances between each sequence in the alignment. This was achieved using the Dayhoff PAM 001 distance matrix (Dayhoff *et al.*, 1978), which generates a transition probability matrix that allows prediction of the probability of changing from one amino acid to any other. These probabilities are scaled in terms of a unit, which is an expected 1% change between two amino acid sequences. The program then uses these probabilities to compute the distance between each pair of taxa. The distance calculated is therefore scaled in units of the expected fraction of amino acids changed. The phylogenetic relationships of the taxa were then recreated from this distance-matrix using the neighbor-joining algorithm (Saitou and Nei, 1987) of the NEIGHBOR program. This approach modifies the raw data of the distance-matrix by adjusting the distances between each pair of taxa on the basis of the average divergence from all other taxa pairs. The algorithm starts by linking the least distant pair of taxa as defined by the modified matrix. Once this has been done the common node, defined by the common ancestor of these two taxa, is added to the tree but the branches are removed. This process converts the common ancestor into a terminal node on a tree of reduced size. This process is repeated, with the loss of two branches at each step, until the process is completed, when there are two nodes separated by a common branch. Though this approach is distinct from parsimony approaches, it still utilises the same principle of maximum parsimony (or minimum-evolution) when generating a tree. However, this method does not necessarily produce the minimum-evolution tree, since, like heuristic methods, it does not generate all possible trees and then select the shortest tree.

It is important to note that the phylogenetic trees generated by these two approaches have very different appearances. Both approaches recreate phylogenetic relationships by linking more closely related sequences by fewer internal nodes. However, unlike the PROTPARS program, the NEIGHBOR program also computes branch lengths based on the degree of similarity of sequences linked by a common node. Thus, the trees generated by the NEIGHBOR program have branch lengths of different sizes proportional to the degree of similarity between sequences while the PROTPARS program generates trees with all branch lengths set to the same size.

4.2.8.1. The predicted proteins encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* cluster with known cathepsin B enzymes

As discussed above, the primary amino acid sequence data suggested that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encoded cathepsin B-like enzymes. In order to confirm this, the nucleotide sequences from a variety of cysteine protease genes were obtained from the Genbank and EMBL databases. These sequences were identified by two searches using the STRINGSEARCH program of the GCG package, with the search terms ‘cysteine protease’ and ‘cathepsin’, respectively. Analysis of these sequences revealed that only 62 encoded the complete predicted protein product, and these were selected for initial phylogenetic analysis. In the case of genomic sequences, information supplied with the sequences was used to remove all non-coding regions, such as 5′ and 3′ flanks and introns, to generate complete coding sequences. For sequences derived from cDNA, all 5′ and 3′ UTR sequences were removed to generate the complete coding sequences. The protein sequences for each of the coding nucleotide sequences obtained were subsequently determined using the TRANSLATE program of the GCG package. The sequences were aligned using the Clustal program version V, and phylogenetic relationships recreated using the NEIGHBOR and PROTPARS programs of the PHYLIP package, version 3.5c. The results of these analyses are shown in Figures 4.18 and 4.19.

The NEIGHBOR program (Figure 4.18) clusters all four protein products of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* with the 17 cathepsin B and B-like sequences included in the analysis, strongly supporting the assignment of these four genes as cathepsin B-like. All these cathepsin B-like sequences clustered to a single internal node representing the common cathepsin B ancestor (node A). Furthermore, this node is well separated from a second node representing the common ancestor of all the other cysteine protease enzymes analysed (node B). The degree of separation between nodes A and B suggests that cathepsin B enzymes have diverged substantially from other classes of cysteine protease. The PROTPARS program obtained two equally parsimonious trees. These two trees were virtually identical and therefore only one tree, and the part of the second tree that differed, is shown (Figure 4.19). The PROTPARS program also clusters all the cathepsin B-like sequences to one end of the unrooted tree. However, four of the *C.elegans* cathepsin B-like genes, *cpr-1*, *cpr-3*, *cpr-4* and *cpr-5*, are placed very near sequences not considered as being cathepsin B-like. In particular, *cpr-4* and *cpr-5* are only separated by three internal nodes from a cysteine protease isolated from *Dictyostelium discoides*, yet are separated from *cpr-1* and *cpr-3* by four internal nodes. By comparison, the NEIGHBOR program separated *cpr-4* and *cpr-5* from their nearest non cathepsin B-like sequence by seven internal nodes, and from *cpr-1* and *cpr-3* by only four internal nodes. Therefore, while the PROTPARS program does support, to some extent, the assignment of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as cathepsin B-like, it does not cluster these sequences as closely to the other cathepsin B-like sequences, or as far from non cathepsin B-like sequences as the NEIGHBOR program. The variation of the phylogenetic relationships recreated by these two programs is also manifested in other regions of the trees. Thus, while in most cases both approaches cluster the same groups of sequences together, the relationships between the members of these groups differ between the two trees. These differences are probably accounted for by the different approaches used. As explained earlier, the PROTPARS program uses a heuristic approach to recreate the phylogenetic relationships. Such approaches do not guarantee the generation of the most parsimonious, or minimum-evolution tree. In the case of the PROTPARS program, an additive approach is used when recreating the phylogenetic relationships between sequences, taking the order of the sequences in the alignment as the order for addition to the tree. Therefore, the program can only be expected to generate locally optimal relationships, and these may differ depending on the order the sequences are added. In essence, the program is unable to take a global approach to recreating the phylogenetic relationships and can only generate the most optimal relationships given the current situation (the input order of the sequences). In such situations bootstrap analysis is useful in determining the reliability of trees inferred by such programs. However, the large size of the dataset used negated this approach as it would have proved to be too lengthy an operation.

In contrast, the NEIGHBOR approach first generates a distance matrix, and therefore organises the input sequences on the basis of similarity. Though these distances are then used to recreate phylogenetic relationships using an additive approach, the sequences are added in a logical manner, starting with the most similar sequences first. In this respect the NEIGHBOR program can recreate phylogenetic relationships using a more global approach and is therefore a more appropriate algorithm to use for such a large dataset. Despite the shortcomings of the PROTPARS approach, it was still able to support results obtained using the NEIGHBOR program. Together, these data strongly support the classification of the genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as cathepsin B-like, and indicates that this class of cysteine protease enzyme is highly diverged from other classes of cysteine protease.

4.2.8.2. Phylogenetic comparison of the proteins encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* with cathepsin B-like sequences in the database

A comparison of the percentage similarities and identities of the prepro- forms and predicted mature forms of the enzyme products of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to *cpr-1* and human cathepsin B, suggested that the cathepsin B-like genes of *C.elegans* were as diverged from one another as from human cathepsin B, with the exception of *cpr-4* and *cpr-5* (Table 4.4). The NEIGHBOR and PROTPARS programs were therefore used to infer the phylogenetic relationships of the 21 cathepsin B-like sequences which clustered together in the previous phylogenetic analysis (Section 4.2.8.1).

The 21 cathepsin B-like sequences were realigned using the CLUSTAL program, version V. The NEIGHBOR and PROTPARS programs were used to recreate the phylogenetic relationships between these sequences as described previously. Bootstrap resampling was also performed with the PROTPARS parsimony analysis. A majority-rule consensus tree was obtained from the 20 trees generated by the bootstrap analysis using the CONSENSUS program of the PHYLIP package, version 3.5c (Felsenstein, 1993). These two approaches generated trees of very similar topology (Figure 4.20), with only three minor differences between the two trees. These differences were essentially in the relationship of AC-4 with the other *H.contortus* cathepsin B-like genes, *cpr-6* with *cpr-4* and *cpr-5*, and human cathepsin B with the other mammalian cathepsin B-like genes. The very similar topologies obtained from the NEIGHBOR and PROTPARS programs used strongly suggest that the overall phylogenetic relationships inferred by these two approaches are reliable.

The cathepsin B-like sequences can be divided into several groups based on taxonomic differences; the protist sequences (*Leishmania mexicana*), the plant sequences (*Nicotiana rustica* and *Triticum aestivum*), the vertebrate sequences (*Homo sapiens*, *Mus musculus*, *Bos taurus* and *Gallus gallus*), the insect sequences (*Sarcophaga peregrina*), the trematode sequences (*Schistosoma japonicum* and *Schistosoma mansoni*) and the nematode sequences (*C.elegans*, *H.contortus* and *O.ostertagi*). Bootstrap analysis indicated that sequences from the same taxonomic group were in most cases reliably clustered together (i.e. 100% of the bootstraps recreated these clusters). However the notable exception to this proved to be with the *C.elegans* sequences. Thus, while *cpr-1* and *cpr-3* were reliably clustered together, as were *cpr-4* and *cpr-5*, neither of these two groups were reliably clustered to each other or to *cpr-6*. This is in contrast to the sequences of *H.contortus* and *O.ostertagi* which were all reliably clustered together. The unreliable clustering of these sequences probably accounts for the discrepancy in the relationship of *cpr-6* to *cpr-4* and *cpr-5* observed between the two trees. The inability of bootstrap analysis to reliably group the cathepsin B-like sequences of *C.elegans* may be a result of the high level of divergence observed between these sequences. The phylogenetic relationships inferred using the NEIGHBOR and PROTPARS program suggest that there are three groups of cathepsin B-like enzymes in *C.elegans*. These three groups are represented by *cpr-1* and *cpr-3*, *cpr-4* and *cpr*-5, and lastly, *cpr-6*. The NEIGHBOR program generates long branch lengths between these three groups, suggesting that they are highly diverged from one another. The degree of divergence observed between the three different groups of cathepsin B-like amino acid sequences of *C.elegans* is far greater than the interspecies divergence observed for human cathepsin B and its homologue in the mouse, the chicken and the cow, suggesting that the cathepsin B-like gene family of *C.elegans* may have diverged early in evolution. For the three groups of predicted proteins encoded by the *C.elegans* cathepsin B-like genes, any one member of a group is almost as diverged from the members of other groups as it is from the vertebrate cathepsin B sequences, but is significantly less diverged from members of the same group. This suggests that these three groups may have arisen from gene duplication events which occurred around the time of separation of the nematode and vertebrate ancestral lineages, but that more recent duplication events may have subsequently generated the members comprising each group. However, an important caveat to this observation is that one must assume an equal evolutionary rate when making such comparisons.

4.2.8.3. Gene architecture as a gauge of evolutionary divergence

As mentioned in the previous section, it is necessary to assume that evolution has occurred at an equal rate, when using the degree of divergence between related sequences to infer how ancient the gene duplication events were which gave rise to such sequences. When comparing orthologous genes (genes related by a common ancestor which perform analogous roles in different species), this assumption may be correct. However, such an assumption may not be valid when making comparisons between paralogous sequences. Paralogous sequences represent genes that, while ultimately having a common ancestor, are derived from a gene duplication event rather than a speciation event. After duplication, the paralogues may have assumed different physiological or enzymatic roles and therefore may have been subjected to different evolutionary pressures. These differences in evolutionary pressure may increase the rate at which the coding sequences of two paralogous genes evolve, resulting in highly diverged genes which have actually arisen from more recent duplication events than would be predicted from the degree of divergence of their amino acid sequences. In such circumstances, it would be useful to select a feature which is unlikely to be affected by differing evolutionary pressures to the extent that coding sequences are. In this respect introns may represent a suitable feature, since evolutionary pressures are unlikely to alter the rate at which gene architecture is altered with time. Furthermore, the mutation events which cause intron loss, gain or displacement and lead to altered gene architectures are less likely to generate functional proteins than substitution mutations that give rise to diverged predicted amino acid sequences. Therefore the frequency with which these mutations become fixed during evolution will be somewhat less than substitution mutations that result in changes in the amino acid sequence. Therefore, it is possible to envisage a situation where more recently duplicated paralogous genes may encode highly diverged protein products but share very similar gene architecture. In such situations, the gene architecture will provide important clues as to how recently the genes diverged. For this reason, I decided to study the relative position of introns in cathepsin B and cathepsin B-like genes from different species.

The intron positions of eight cathepsin B and B-like genes were determined from their gene architectures . The predicted amino acid sequences of these genes, and the relative position of their introns, are shown in Figure 4.21. Before discussing the evolutionary information that can be inferred from the intron positions of the eight genes analysed, a brief overview of the patterns of intron distribution in these genes will be mentioned. A comparison of the relative intron positions between the eight genes reveals that a total of six intron sites are conserved between the genes (Figure

4.21, A-F). These sites are not only conserved with respect to the amino acid sequence but also with respect to the exact nucleotide residue at which intron / exon boundaries occur within codons. This level of conservation strongly suggests that these eight genes are derived from a common ancestor. However, there is also evidence to suggest that intron loss, gain and displacement has occurred during the evolution of these genes. If one looks at the intron positions defined by the 11 introns of the *H.contortus* AC-2 gene and *O.ostertagi* CP-1 gene, it becomes apparent that there are several incidences where no introns are found at the equivalent position in the other genes (Figure 4.21, A-H), suggestive of intron loss. A clear example of this is seen for *cpr-6* and mouse cathepsin B at sites E and F respectively. In this example, *cpr-6* has maintained the intron at position E and lost the intron at position F, while the opposite has occurred with mouse cathepsin B. There are also some incidences where the introns of the *C.elegans* cathepsin B-like genes and the mouse cathepsin B gene are found in the vicinity of those sites defined by AC-2 and CP-1, but not at exactly the same positions (Figure 4.21, A, I, J and K), suggesting that intron displacement has occurred with some of these genes presumably as a result of complementary mutation events occurring at each end of the intron. Finally, there are two cases where introns are found in close proximity to each other, but outwith those sites defined by AC-2 and CP-1 (Figure 4.21, L and M). Such positions may have arisen as a result of intron loss in CP-1 and AC-2 followed by intron displacement in the genes containing these introns. Alternatively, a simpler explanation may be that introns were inserted at these sites, after divergence from the *H.contortus* and *O.ostertagi* ancestral lineages. Such complex patterns of intron position have also been observed for other gene families. Indeed, the nemoglobin family of diverged nematode globins shows similar types of variation in intron position which have also been explained by models of intron loss, gain and displacement (Blaxter, 1993).

For evolutionary analysis, the AC-2 and CP-1 genes of the parasitic nematode species *H.contortus* and *O.ostertagi* provide a useful baseline for comparison with the other genes. Phylogenetic analysis of the protein products of AC-2 and CP-1 (Figure 4.20) suggest that these two genes are very closely related and this is reflected by their very similar gene architectures. Thus, both genes possess 11 introns and the position of all these introns are not only conserved with respect to the amino acid sequence but also to the exact nucleotides at which intron / exon boundaries occur within codons (Figure

4.21). Furthermore, these two genes define all the common intron sites for the genes analysed since the *C.elegans* cathepsin B-like genes and the mouse cathepsin B gene only share common intron positions at sites already determined by AC-2 and CP-1 (Figure 4.21, A-F). Accordingly, the 11 intron positions defined by AC-2 and CP-1 were used as the basis of comparison for the five *C.elegans* cathepsin B-like genes and the mouse cathepsin B gene included in this analysis.

Alignment of the amino acid sequences of the eight genes revealed that only six of the 11 common intron positions defined by AC-2 and CP-1 are shared with the five *C.elegans* cathepsin B-like genes and the mouse cathepsin B gene (Figure 4.21, A-F). This suggests that the *C.elegans* and mouse genes are substantially more diverged from AC-2 and CP-1 than these two genes are from one another. This conclusion is also supported by the variation in intron number of the *C.elegans* and mouse genes. All six of these genes possess substantially fewer introns in their coding region than AC-2 and CP-1, with *cpr-6* and mouse cathepsin B having the most (seven introns) and *cpr-4* the least (one intron). Furthermore, none of the *C.elegans* and mouse genes each share more than four common intron positions with AC-2 and CP-1. Thus, while mouse cathepsin B and *cpr-6* both possess seven introns in their coding regions, each shares only four and three common intron positions, respectively, with those sites defined by AC-2 and CP-1. Together these data clearly suggest that the five *C.elegans* cathepsin B-like genes and the mouse cathepsin B gene are all much more diverged from AC-2 and CP-1 than these two genes are from one another. This conclusion was also drawn from the phylogenetic analysis involving the predicted proteins of these eight genes (Figure 4.20) and therefore supports the efficacy of using gene architecture, as defined by intron position, as a means of gauging evolutionary divergence.

A comparison of common intron positions between the eight genes (Figure 4.21, A-F) strongly suggests that most of the *C.elegans* cathepsin B-like genes are as diverged from one another as they are from mouse cathepsin B. None of the *C.elegans* cathepsin B-like genes share more common intron positions with one another than they do with mouse cathepsin B. In fact, no more than one common intron position is shared between any of the *C.elegans* and mouse genes. For example *cpr-6* shares only one common site with *cpr-1*, *cpr-4* and the mouse cathepsin B gene, and no common sites with *cpr-3* or *cpr-5*. This level of divergence suggests that the *C.elegans* cathepsin B-like genes have arisen from ancient gene duplication events.

Phylogenetic analysis of the predicted proteins encoded by the five cathepsin B-like genes of *C.elegans* (Figure, 4.20) clusters these sequences into three groups and suggests that each group may have arisen from gene duplication events which occurred as early as the time of separation of the nematode and vertebrate ancestral lineages, but that the members within each group may have arisen from more recent duplication events. The three groups comprise; *cpr-1* and *cpr-3*, *cpr-4* and *cpr-5*, and *cpr-6*. In some cases, the results obtained from phylogenetic analysis of the predicted encoded proteins agree with those obtained from analysing the gene architectures. For example, phylogenetic analysis of the predicted encoded proteins suggests that the group defined by *cpr-6* is highly diverged from the other two groups. This is supported by the gene architectures, as defined by intron position and number. *cpr-6* possesses seven introns while the members of the other two groups possess only between one and three introns. Furthermore, *cpr-6* does not share more than one intron position in common with any of the four other *C.elegans* cathepsin B-like genes comprising the other two groups. Phylogenetic analysis suggests that *cpr-1* and *cpr-3* may have arisen from more recent gene duplication events than those that generated the three groups. This is supported, at least in part, by comparison of the gene architectures of these two genes. *cpr-1* and *cpr-3* have a similar number of introns, two and three introns respectively. Furthermore, intron I of *cpr-1* is situated in close proximity to intron I of *cpr-3* (Figure 4.21, A), and intron II of *cpr-1* is situated at an identical position to intron II of *cpr-3* (Figure 4.21, D). Though these two genes do possess some similarities in their gene architectures (suggestive of a more recent gene duplication event) their gene architectures are sufficiently diverged to suggest that this duplication event may have occurred soon after the events predicted to have generated the three groups. This is supported by phylogenetic analysis using the NEIGHBOR program which, links the products of these two genes by a common node but separates them with long branch lengths (Figure 20C).

For *cpr-4* and *cpr-5*, there is no such correlation between the similarity of their predicted proteins and their gene architectures. The putative enzymes encoded by *cpr-4* and *cpr-5* show higher levels of homology to each other than the predicted proteins encoded by any other pair of *C.elegans* cathepsin B-like genes (Table 4.4) and this is reflected by the results obtained from phylogenetic analysis (Figure 4.20C). However, their gene architectures, as defined by intron number and position, are very different. *cpr-4* possesses a single intron and the position of this intron is not conserved with either of the two introns of *cpr-5*, nor are the introns of these two genes found in close proximity to one another. This suggests that *cpr-4* and *cpr-5* may have arisen from a much more ancient gene duplication event than expected from the similarity of their protein sequences. Accordingly, the coding regions of these two genes may have evolved either in parallel or at a much slower rate after gene duplication than the other *C.elegans* genes, suggesting that these two genes may encode enzymes which are required for similar biological functions.

4.3. Discussion

4.3.1. A cathepsin B-like multigene family in *C.elegans*

The results presented in this chapter indicate that *C.elegans* possesses a cathepsin B-like cysteine protease multigene family comprising the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (Chapter 3) and the previously isolated cathepsin B-like gene, *cpr-1* (Ray and McKerrow, 1992). Sequence analysis indicates that the putative cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* possess all the features expected for this class of enzyme (Section 4.2.7). Furthermore, phylogenetic analysis clusters the predicted enzyme products of these four genes with other cathepsin B sequences in the database and away from other cysteine protease classes (Section 4.2.8.1).

4.3.2. The Implications of isolating a cathepsin B-like multigene family from *C.elegans*

Until now, cathepsin B-like multigene families had only been reported in parasitic nematode and trematode species (Chapter 1, Section 1.9) leading to suggestions that amplification of this gene family may have been important for the parasitism of these species (Pratt *et al.*, 1992a). However, the presence of a cathepsin B-like multigene family in the free living nematode *C.elegans*, indicates that such gene families are not unique to parasitic nematode and trematode species *per se*.

4.3.2.1. *C.elegans*, *H.contortus* and *O.ostertagi* may possess more, as yet, unidentified cathepsin B-like genes

Phylogenetic comparisons using the NEIGHBOR and PROTPARS programs of the PHYLIP package, reveal interesting differences in the relationships of the predicted products of the cathepsin B-like genes of *C.elegans*, both to each other and to the products of cathepsin B, and cathepsin B-like genes from other organisms (Section 4.2.8.2). In particular, the products of the cathepsin B-like genes of *C.elegans* are much more diverged from each other than those isolated from *H.contortus* and *O.ostertagi*. This may be explained by the methods used to isolate these genes. The *H.contortus* cathepsin B-like genes AC-2, AC-3 and AC-4 and the *O.ostertagi* gene CP-1 were all isolated by homology to the previously isolated gene AC-1. AC-1 was initially isolated by screening a cDNA expression library using antisera generated from proteins purified from anticoagulant extracts of *H.contortus* (Cox *et al.*, 1990). AC-2 was subsequently isolated from an *H.contortus* genomic DNA library using AC-1 as a probe (Pratt *et al.*, 1990). These two sequences share 97% nucleotide sequence identity and 98% amino acid sequence identity in their coding regions. Though it is unclear whether these sequences represent distinct genes, or population polymorphisms, the authors argue that the number of differences is more consistent with their being distinct genes. AC-3 was initially identified by low stringency hybridisation to the genomic clone containing AC-2 using a 40 nucleotide oligomer corresponding to the 5′ end of the AC-1 cDNA , also present in AC-2 (Pratt *et al.*, 1992a). Subsequent sequence analysis allowed the design of a non-degenerate oligonucleotide primer which was used, in conjunction with an oligo (dT) primer, to amplify sequences from *H.contortus* cDNA, and yielded cDNA clones for AC-3 and AC-4. CP-1 and CP-2 (a partial genomic sequence) were isolated by screening an *O.ostertagi* genomic DNA library with the AC-1 cDNA of *H.contortus* (Pratt *et al.*, 1992b). Such homology based methods inevitably bias the screen by failing to detect divergent members.

In contrast, the method by which the genes reported here were identified has no such bias since the cDNA clones used to isolate *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were originally identified by partial sequencing of randomly selected clones (Waterston *et al.*, 1992). Indeed, the library from which the cDNA clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* were obtained, may have biased against the isolation of highly homologous clones because the library was partially normalised prior to isolation of the cDNA clones. This normalisation was achieved by screening a lambda SHLX2 *C.elegans* cDNA library with cDNA clones and selecting only clones that did not hybridise. These clones were then added to those used for probing and the process repeated several times (Palazzolo *et al.*, 1990).

The differences in the degree of similarity of the genes cloned by these two approaches may therefore be due to different biases imposed by the methods used for their isolation. It is therefore possible that these nematode species may possess more members of the cathepsin B-like gene family than those so far identified. If this is the case, one might expect the cathepsin B-like multigene families of nematode species to be organised into several distinct highly diverged groups (as suggested from the *C.elegans* data), with each group comprising several highly homologous members (as suggested by the *H.contortus* and *O.ostertagi* data).

4.3.2.2. The cathepsin B-like multigene families of free living and parasitic nematode species may have arisen from different gene duplication events

The greater degree of divergence observed between most of the *C.elegans* cathepsin B-like genes than that observed between the *H.contortus* and *O.ostertagi* cathepsin B-like genes may reflect a genuine difference between these two species, not just different biases in the methods used to isolate these genes. In such a situation some of the members of the cathepsin B-like multigene families of *H.contortus* and *O.ostertagi* may have arisen from more recent duplication events than those of *C.elegans*, as a result of evolutionary pressures unique to these parasitic nematode species.

Most of the cathepsin B-like genes of *C.elegans* are sufficiently diverged to suggest that they arose from duplication events which occurred early in evolution, possibly as early as the time of separation of the nematode and vertebrate ancestral lineages (Sections 4.2.8.2 and 4.2.8.3). Since *H.contortus*, *O.ostertagi* and *C.elegans* do not represent highly diverged taxa, being of the same Subclass (Secerentea), the ancestral lineages of these nematode species probably separated after the ancient gene duplication events predicted to have generated the three groups of cathepsin B-like genes in *C.elegans*. Accordingly, these three nematode species are likely to possess cathepsin B-like genes which arose from the same duplication events that generated the three groups of cathepsin B-like genes identified in *C.elegans*. However, subsequent duplication events in the parasitic nematode species, after separation of the *C.elegans* ancestral lineage, may have generated additional members of this multigene family in these parasitic nematode species. This scenario predicts that *H.contortus* and *O.ostertagi* may possess more cathepsin B-like genes than *C.elegans* and that these additional genes will be very closely related. There are several pieces of evidence to suggest that this is indeed the case. There is circumstantial evidence to suggest that *H.contortus* may possess more cathepsin B-like genes than *C.elegans*. At least six additional cathepsin B-like genes have been identified from *H.contortus* (D.Knox, pers. comm.) indicating that, in this parasitic nematode species, the gene family comprises at least 11 members. In contrast, only one more cathepsin B-like cDNA clone (distinct from the cathepsin B-like genes isolated to date) has been isolated by partial sequencing of randomly isolated cDNA clones from *C.elegans* (McCombie *et al.*, 1992) and no additional cathepsin B-like genes have been identified in the 22.5 Mb of *C.elegans* genomic DNA sequenced to date. There is also evidence to suggest that the *H.contortus* and *O.ostertagi* cathepsin B-like genes characterised to date have arisen from more recent duplication events than the *C.elegans* cathepsin B-like genes. The gene architectures (as defined by intron number and position) of AC-2 and CP-1 from *H.contortus* and *O.ostertagi* are identical, both possessing 11 introns at identical positions. Partial sequencing of the 3′ end of the *O.ostertagi* CP-3 gene has identified four introns which are in identical positions to introns 9 - 11 of CP-1 and AC-2 (Pratt *et al.*, 1992b). In addition, partial sequencing of the 5′ end of the *H.contortus* AC-3 gene has identified four introns which are in identical positions to introns 1 - 4 of AC-2 and CP-1 (Pratt *et al.*, 1992a). These results suggest that the four genes, AC-2 and AC-3 from *H.contortus* and CP-1 and CP-3 from *O.ostertagi*, may all have introns in identical positions and therefore share very similar gene architectures. These observations suggest that these genes have arisen from a much more recent duplication event than the *C.elegans* cathepsin B-like genes. One would expect the ancestor of these duplication events to be one of the genes generated by the same ancient duplication event predicted to have generated the three groups of *C.elegans* cathepsin B-like genes. Accordingly, one might expect to find at least one cathepsin B-like gene in *C.elegans* that is significantly more closely related to the cathepsin B-like genes isolated from *H.contortus* and *O.ostertagi*, which would represent the orthologue of the gene amplified in these parasitic nematode species. Such a gene has not as yet been identified in *C.elegans* since none of the cathepsin B-like genes isolated from *C.elegans* to date possess gene architectures or encode enzymes which are significantly more similar to the *H.contortus* and *O.ostertagi* genes. Therefore, there may be at least one, as yet uncharacterised, cathepsin B-like gene in the genome of *C.elegans* which will show more homology to the AC and CP cathepsin B-like genes of *H.contortus* and *O.ostertagi* than the genes isolated to date.

If the members of the cathepsin B-like multigene families of *H.contortus* and *O.ostertagi* isolated to date have arisen in the manner suggested above, the enzymes encoded by these genes may have important roles required for parasitism by these two species. Accordingly, although the isolation of a cathepsin B-like multigene family from the free living nematode *C.elegans* indicates that such multigene families are not unique to parasitic nematode species *per se*; it may not disprove the specific function of some cathepsin B-like genes in parasitic processes.

4.3.2.3. Cathepsin B-like multigene families may be present in other metazoâns

In order to argue the presence of more diverged cathepsin B-like genes in the genomes of higher eukaryotes, or indeed of *H.contortus* or *O.ostertagi*, one must identify examples of genes which encode cathepsin B-like enzymes that are sufficiently diverged not to cross hybridise with one another. The COMPARE and GAP analyses performed between the genomic sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* indicate these genes are highly diverged (Section 4.2.5), with only short regions of homology occurring between the genes in the regions which encode the catalytic residues, C, H and N. This suggests that these genes are sufficiently diverged such that detection of one, using another as a hybridisation probe, would be extremely difficult, or impossible. The degree of divergence therefore suggests that it is possible that more diverged cathepsin B-like genes might not be detected with a homology based screen, and therefore that

H.contortus and *O.ostertagi* may possess yet more members of this gene family than those isolated to date. This observation has wider implications since additional diverged cathepsin B-like genes may also exist in other metazoans in which only single cathepsin B genes have so far been found.

For example, there is a large amount of circumstantial evidence to suggest that the human and mouse cathepsin B genes are single copy. The evidence gathered to date which suggests that the human and mouse cathepsin B genes are single copy, comes from data obtained during library screens to obtain cDNA clones and genomic clones for these genes. Thus, the first cDNA clones of human and mouse cathepsin B to be identified from lambda gt11 library screens were all overlapping (Chan *et al.,* 1986). The authors concluded that each of these genes are single copy. This conclusion was supported by the simple hybridisation patterns obtained by hybridising Southern blots of restriction endonuclease digested human genomic DNA with a human cathepsin B cDNA clone (Fong *et al.,* 1986). The genes encoding human and mouse cathepsin B have subsequently been cloned (Gong *et al.,* 1993; Ferrara *et al.,* 1990; Qian *et al.,* 1991). More recently, a number of human cathepsin B cDNA clones have been cloned, sequenced (Gong *et al.,* 1993; Cao *et al.,* 1994; Tam *et al.,* 1994) and compared to the human cathepsin B gene sequence. These comparisons all suggest that the cDNA clones isolated are all derived from the same human cathepsin B gene. These and other studies have resulted in the current opinion that human and mouse cathepsin B (and therefore probably also the rat and bovine cathepsin B homologues) are present as single copies in their respective genomes. This does not dispute the presence of additional, diverged cathepsin B-like genes within the human and mouse genomes, since such genes may have remained undetected by the homology based screens that have been used to obtain evidence for the single copy nature of the human and mouse cathepsin B genes. Thus there may also be additional, diverged cathepsin B-like genes present in the human and mouse genomes which have not as yet been detected. This would be very similar to the situation in *C.elegans*, where the results suggest that each of the cathepsin B-like genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, are single copy (Chapter 3) and are sufficiently diverged such that detection of one using another as a hybridisation probe may be difficult or impossible.

Alternatively, the existence of a cathepsin B-like multigene family within nematode species may reflect functions specific to nematode biology. Accordingly, vertebrates may not possess additional, diverged cathepsin B-like enzymes. If this proves to be the case, cathepsin B-like enzymes may prove to be very useful targets for anthelminthic agents in the treatment of parasitic nematode diseases.

4.3.3. The cathepsin B-like genes of *C.elegans* may encode enzymes with different biological functions

As stated above, the predicted amino acid sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* strongly suggest that these genes encode cathepsin B-like enzymes. However, these putative enzymes are much more diverged from one another than those encoded by the genes, AC-1, AC-2, AC-3, AC-4 and AC-5 of *H.contortus*. The five cathepsin B-like genes of *C.elegans* can be divided into three groups, on the basis of the similarity of their predicted enzyme products (see Section 4.2.8.2). Phylogenetic analysis reveals that each of these three groups are almost as diverged from one another as they are from the vertebrate cathepsin B enzymes. Such a high degree of intraspecies divergence suggests that the three groups of enzymes may have evolved independently to assume distinct biological functions.

4.3.3.1. *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may encode enzymes with different specificities and activities

Recombinant rat cathepsin B has been expressed in yeast and active enzyme has accumulated in the media under optimal culture conditions (Lee *et al.*, 1990). Detailed kinetic analysis of the purified recombinant protein obtained indicates that enzymatically it is essentially identical to rat liver cathepsin B (Hasnain *et al.*, 1992). This system has allowed site-directed mutagenesis approaches to be used to study the effect of altering several residues previously implicated as being important in determining the activity of cathepsin B. These studies (Hasnian *et al.*, 1992) have identified two regions outside the catalytic residues C, H and N, which affect the substrate specificity and activity of rat cathepsin B. Cathepsin B appears to be unusual amongst other members of the cysteine protease group of proteolytic enzymes in its ability to function as both an endopeptidase and a peptidyldipeptidase (Aronson and Barrett, 1978). Two adjacent histidine residues at position 110 and 111 of rat cathepsin B are thought to play an important role in this peptidyldipeptidase activity (exopeptidase activity). Mutation of one of the histidine residues, His(111), to glutamine results in a much reduced exopeptidase activity and increased endopeptidase activity with respect to wild-type rat cathepsin B, providing strong evidence that the dihistidine repeat plays an important role in exopeptidase activity. In the second region, mutation of the glutamic acid residue (position 245 in mature rat cathepsin B) to glutamine or alanine alters the substrate specificity of the recombinant enzyme. This residue, Glu(245), is thought to be responsible for vertebrate cathepsin B's characteristic activity against synthetic substrates possessing two basic amino acid residues. Accordingly, mutation of this residue in recombinant rat cathepsin B drastically reduces the rate at which this enzyme cleaves the synthetic substrate Z-Arg-Arg-pNA, compared to wild type.

The equivalent residues for human cathepsin B are shown in bold and boxed (Section 4.2.7, Figure 4.8). A comparison of these residues to residues at the equivalent position in *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* reveals interesting differences between these enzymes at both these sites. Only *cpr-6* contains the adjacent histidine residues involved in exopeptidase activity. *cpr-4* and *cpr*-5 both share the same non-conservative substitutions of glutamic acid and threonine at this site while *cpr-1* and *cpr-3* share a common seven amino acid deletion of this region. This suggests that *cpr-1*, *cpr-3*, *cpr-4* and *cpr-5* may have reduced exopeptidase activity with respect to both *cpr-6* and human cathepsin B. The predicted enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* can be grouped into three classes on the basis of; presence of the dihistidine repeat, substitution of the dihistidine repeat, or deletion of the dihistidine repeat. It is interesting to note that these three groups correlate with the three groups defined by phylogenetic analysis (Section 4.2.8.2). At the second site, equivalent to Glu(245) of rat cathepsin B, none of the *C.elegans* cathepsin B-like enzymes possess this glutamic acid residue, suggesting that these enzymes may not possess significant activities against synthetic substrates with two basic amino acid residues. Furthermore, the substitutions made at this site are not conservative. These data suggest the cathepsin B-like enzymes isolated from *C.elegans* may have different specificities and activities not only from mammalian cathepsin B, but also from each other.

4.3.3.2. The enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be subjected to different types of post-translational modification

Only *cpr-3*, *cpr-4* and *cpr-6* encode products with potential N-glycosylation sites (Asn-X-Ser/Thr) and their positions are not conserved. Studies with mammalian cells have demonstrated that glycosylation is an integral part of lysosomal trafficking, with mannose-6-phosphate being the recognition signal for transport of soluble proteins to the lysosomes (von Figura and Hasilik, 1986). Briefly, N-linked glycosylation of certain asparagine residues of proteins occurs in the lumen of the endoplasmic reticulum and these glycosylated proteins are subsequently transported to the Golgi complex. In the Golgi complex, mannose residues of lysosomal proteins are phosphorylated. The specificity of this phosphorylation step is thought to be generated by a signal within the lysosomal protein. This signal is sensitive to treatment with heat, SDS and trypsin (Lang *et al.*, 1984) and is therefore thought to be dependent on the tertiary rather than primary structure of the lysosomal protein. Thus, the generation of a mannose-6-phosphate lysosomal targeting signal in the Golgi complex requires the initial presence of a glycosylated asparagine residue. Treatment with tunicamycin, which inhibits protein glycosylation, results in most cell lines secreting proteins which are normally directed to the lysosomes (Rosenfeld *et al.*, 1982). These data suggest that the gene products of *cpr-1* and *cpr-5*, which both lack potential N-glycosylation sites, may be secreted as opposed to being directed to the lysosomes. Alternatively, it is possible that the enzymes encoded by these two genes may be directed to the lysosomes by a mannose-6-phosphate independent pathway. There is evidence to suggest that such carbohydrate-independent mechanisms do exist. This conclusion has been based primarily on work with hepatocytes from I-cell diseased patients (Kornfeld, 1986). I-cell disease is a disorder which affects mannose-6-phosphate biosynthesis. In some patients suffering from this disease, lysosomal proteins are still delivered to the lysosomes, despite the absence of a mannose-6-phosphate marker. Other studies, with HepG2 cultured cells treated with tunicamycin, have also reported targeting of lysosomal proteins to the lysosomes (Rijnboutt *et al.,* 1991).

Phylogenetic analysis clustered the five cathepsin B-like enzymes of *C.elegans* into three groups (Section 4.2.8.2); group 1, *cpr-1* and *cpr-3*; group 2, *cpr-4* and *cpr-5*; group 3, *cpr-6*. Only one enzyme in each group possesses a potential N-glycosylation signal. This suggests that the more similar enzymes within a group may have functionally distinct roles, possibly with one being secreted and one being retained within the lysosomes of the cells which express them.

Finally, both *cpr-3* and *cpr-6* encode C-terminal extensions (31 and 19 amino acids respectively) from the site of cleavage of the human C-terminal peptide (Chan *et al.,* 1986). In contrast, the predicted protein sequences of *cpr-1*, *cpr-4* and *cpr-5* all end in the region of the predicted site of cleavage of the human C-terminal peptide, suggesting that these enzymes do not possess such a peptide and therefore lack this aspect of processing. It is not known whether the C-terminal extensions encoded by *cpr-3* and *cpr-6* are present in the mature enzymes, as previously reported for *Trypanosoma cruzi* (Cazzulo *et al.,* 1992), or represent much larger C-terminal peptides which are cleaved. Though the role of these extensions is unknown, their presence is likely to reflect differences in either enzyme activity or substrate specificity, or differences in post-translational processing.

Interestingly, a single intron is present in the region of each of the *cpr-3* and *cpr-6* genes which encode the C-terminal extensions. The introns-early hypothesis proposes that spliceosomal introns predate the divergence of prokaryotes and eukaryotes and that each exon encodes some 'functional unit' of a protein (Doolittle and Stoltzfus, 1993). This hypothesis also suggests that novel proteins can be generated by shuffling these exons around during evolution. Therefore, the presence of introns in the C-terminal extensions of *cpr-3* and *cpr-6*, may define some additional functional domain present in these two enzymes.

4.3.4. Regulation of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The 5′ RACE experiments generated RACE clones for each gene whose 5′ ends cluster to a single region of the 5′ flank of each gene (Section 4.2.6), suggesting that each gene is under the control of a single promoter region. Northern analysis (Chapter 5, Section 5.2.2) later revealed that the predominant transcripts from each of these four genes are of a single size species. Furthermore, no evidence for alternative splicing was observed for these four genes within the limits of resolution and detection of Northern analysis. This is in complete contrast to the human and mouse cathepsin B genes, where the data suggest that these two genes are both transcribed from multiple promoters and that their transcripts also undergo alternative splicing.

Northern analysis of RNA isolated from several normal murine tissues, as well as murine tumours, have revealed multiple mRNA transcript species for mouse cathepsin B (Qian *et al.,* 1989; Moin *et al.,* 1989). Similar experiments with normal human tissues and human tumours have revealed a similar situation for human cathepsin B (Page *et al.,* 1992; Gong *et al.,* 1993; Tam *et al.,* 1994). The cloning and characterisation of human cathepsin B (Gong *et al.,* 1993) and mouse cathepsin B (Ferrara *et al.,* 1990; Qian *et al.,* 1991) allows the comparison of cDNA clones, generated from the different mRNA species of these two genes, with the appropriate cloned gene. As a result of this, a number of cDNA clones derived from different sized transcripts from the human cathepsin B gene have been cloned and sequenced (Gong *et al.,* 1993). Comparison of these transcripts to cloned human cathepsin B suggests that the different transcripts arose by alternative splicing of a single gene. Five different species of transcript were identified. Four of these contain identical coding regions, differing only in their 5' and 3′ UTRs. The fifth lacks the signal prepeptide and a portion of the proregion, as a result of skipping exons two and three. However, this transcript species was only identified in human breast and colon carcinomas and in a human melanoma, not in normal tissues, and may therefore represent an aberrant transcript. The identification of at least four transcript species with identical coding regions but varied 5′ and 3′ ends suggest that human cathepsin B expression may be regulated in part at the level of RNA processing. Interestingly, two new exons of the human cathepsin B gene have recently been identified in the 5′ UTR between exons two and three, and have been named 2a and 2b (Berquin, I *et al.,* 1995). Reverse Transcription-Polymerase Chain Reaction (RT-PCR) and primer extension assays performed with matched normal / tumour samples have demonstrated that cathepsin B mRNA species vary both between different cell types and between different differentiation states, and that there is a trend towards preferential expression of exon 2a by tumours. These results suggest that RNA processing may perform an important role in the regulation of human cathepsin B in different tissues.

The same RT-PCR and primer extension assays (Berquin, I *et al.,* 1995) have also demonstrated that transcription of human cathepsin B is initiated from more than one promoter region. This may allow for differential regulation of human cathepsin B expression from different promoters. A RACE approach has also been used to identify at least three different leader sequences for murine cathepsin B (Rhaissi *et al.,* 1993). By comparing the nucleotide sequences of the RACE products to the 5′ flank region of mouse cathepsin B, the authors confirmed that the RACE products were all derived from the same transcriptional unit but that transcription was being initiated at more than one promoter region. This suggests a similar mode of regulation for mouse cathepsin B as for human cathepsin B.

The added level of complexity in the regulation of human and mouse cathepsin B expression may reflect the greater complexity of these vertebrate species when compared to *C.elegans*. Thus, cathepsin B expression in vertebrates may have to be more tightly regulated than cathepsin B-like expression in *C.elegans*. Alternatively, the differences in the regulation of human and mouse cathepsin B gene expression, when compared to the regulation of *C.elegans cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* cathepsin B-like gene expression, may reflect different approaches to ensuring appropriate expression of cathepsin B. Thus vertebrates may utilise a complex regulatory system to differentially regulate expression of a single cathepsin B gene in different tissues while *C.elegans* may use several genes which can each be regulated independently from one another to ensure expression of the appropriate gene, or combinations of genes, in the appropriate tissue and/or at the appropriate time. Such differences in the mode of regulation may also reflect differences in the biological roles of these enzymes between vertebrate and nematode species.

4.3.4.1. Post-transcriptional regulation

Proteins which bind RNA secondary structure in the 5′ and 3′ UTR sequence of mature mRNA transcripts have been shown to be important for regulating the translation and stability of those transcripts. A good example of this is the iron-response element involved in regulating intracellular iron concentration. Translation of the ferritin iron-binding protein is thought to be blocked by a regulatory protein which binds an iron-response element in the 5′ UTR of the mature transcript of ferritin. A similar iron response element to that found in the 5′ UTR of the ferritin mRNA is also present in the 3′ UTR of the transferrin receptor, responsible for uptake of iron in eukaryotic cells. Binding of the same regulatory protein is thought to stabilise the mature mRNA of the transferrin receptor, by blocking degradation of this mRNA (Casey *et al.,* 1988).

The mature mRNA transcript of *cpr-3* (1,329 bases, not including a poly(A) tail) is the largest generated of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. Since, *cpr-3* does not encode the largest protein product, this increase in size is a result of a larger 3′ UTR in the *cpr-3* transcript. The FOLDRNA program generates the most stable stem-loop structure with this 3′ UTR (-20.1 kcal at 37°C). The presence of a stable stem-loop structure alone cannot be considered indicative of its functional role. However, the presence of such a structure makes post-transcriptional regulation of *cpr-3* expression possible. In support of this, two cathepsin B-encoding cDNAs, *hCBF1* and *hCBF2*, have been isolated from a normal human embryonic fibroblast library. These cDNA clones show differences in their 3′ UTRs from the published sequences of the human hepatoma and kidney cathepsin B-encoding cDNAs. Both *hCBF1* and *hCBF2* possess a 10 bp insertion in the 3′ UTR that may allow the formation of a highly stable stem-loop structure, absent in mRNA species without this insertion (Tam *et al.,* 1994). The authors indicate that the introduction of this 10 bp insertion generates a predicted stem-loop structure with a potential free energy of -17.2 kcal, by comparison to the -8.8 kcal stem-loop structure of cDNAs which do not contain this insertion. The authors suggest that the increased stability of the secondary structure of the 3′ UTR may affect the stability of this transcript. These data provide evidence that expression of cathepsin B may be regulated post-transcriptionally in higher eukaryotes. Therefore, it is possible that the transcript of *cpr-3* may also be regulated by an analogous system. However, it should be noted from the discussion in Section 4.3.4, that regulation of vertebrate cathepsin B gene expression may be very different to regulation of *C.elegans* cathepsin B-like gene expression, and therefore that regulation of cathepsin B at the level of transcript stability may be specific to vertebrates

4.3.4.2. Expression of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be temporally and/or spatially restricted

Sequence analysis of the 5′ flank regions of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* reveals clustering of several putative *cis*-acting regulatory elements upstream of the region of transcription initiation, determined by 5′ RACE (Section 4.2.6, Figure 4.17). For *cpr-4* and *cpr-6*, putative TATA box (TATAWAW) promoter elements were identified in the correct position with respect to transcription initiation. For *cpr-3* and *cpr-5*, no such potential TATA box promoter sequences were identified indicating that a TATA-less promoter must be involved in transcription of these two genes. For all four genes, between two and five classic GATA motifs (WGATAR) were identified clustered in the 5′ flank regions, plus additional related variants which might also bind the GATA family of transcription factors (Ko and Engel, 1993; Merika and Orkin, 1993).

GATA factors are widely recognised as being important regulators of a variety of lineage restricted genes in vertebrates. The first GATA factor, GATA-1, was originally identified as an abundant erythroid cell-specific protein that bound the WGATAR consensus sequence found in the regulatory region of several globin and non-globin erythroid specific genes (Wall *et al.,* 1988; Evans *et al.,* 1988; Knezetic and Felsenfeld, 1989). After cloning of GATA-1 (Tsai *et al.,* 1989), three further GATA factors were subsequently characterised, GATA-2, GATA-3 and GATA-4 (Orkin, 1992). All four GATA factors were found to share a high degree of amino acid identity in their two zinc-finger DNA-binding domains. Binding site enrichment studies have shown that each of these GATA factors have distinct but overlapping preferences for different GATA motifs (Ko and Engel, 1993; Merika and Orkin, 1993).

The four GATA factors all have distinct but overlapping patterns of tissue specific expression. For example, GATA-1 is expressed by cells of the erythroid lineage and eosinophils (Zon *et al.,* 1993) while GATA-2 is expressed in neutrophils and in the developing brain (Orkin, 1992), but both are expressed in mast cells and the megakaryocytic lineages.

The *cis*-acting GATA motif, to which GATA proteins bind, was first identified in studies of the globin genes. Numerous other erythroid genes have subsequently been isolated which require GATA elements for correct expression. Furthermore, a response element which directs erythroid-restricted expression has been generated by combining a GATA motif with a CACCC motif upstream of a heterologous promoter (Walters and Martin, 1992).

Functionally important GATA motifs have been found in several different environments. They have been found in both promoter elements (either upstream of, or replacing, the TATA box) and within enhancer elements (Orkin, 1992). GATA factors may well perform different functions in these different settings. For example, GATA factors may inhibit expression of some genes by interacting with their core promoters and preventing formation of the pre initiation complex by steric hindrance. The core promoters of the rat platelet factor 4 (PF4) and mouse erythropoietin genes both possess GATA motifs in place of a TATA box. Initiation of transcription from these promoters requires the TATA-binding protein of TFIID and is inhibited by GATA-1 in *in vitro* transcription assays (Aird *et al.,* 1994). In addition, using mobility shift assays, the authors have shown that GATA-1 interacts with the core promoter GATA motif and prevents formation of the preinitiation complex. In other settings, the binding of GATA proteins to sequences upstream of the core promoter may enhance transcription by promoting formation of the pre initiation complex (Ravid *et al.,* 1991).

Putative GATA regulatory motifs have been implicated in the regulation of several *C.elegans* genes which show spatially restricted patterns of expression including; *cpr-1* (Ray and McKerrow, 1992), *dpy-7* (J.Gilleard, pers. comm.), the six vitellogenin genes (Spieth *et al.,* 1985) including *vit-2* (MacMorris *et al.,* 1992), and the gut esterase gene, *ges-1* (Stroeher *et al.,* 1994). Indeed , protein factors from nuclear extracts of *C.elegans* embryos have been shown to bind the GATA motifs of *ges-1* (Stroeher *et al.,* 1994).

Several GATA factors have also been identified in *C.elegans*; *elt-1* (Spieth *et al.,* 1991), *elt-2* (Hawkins and McGhee, 1995), *elt-3* (J.Gilleard, pers. comm.) and *end-1* (J.Zhu, pers. comm.). *elt-1* was originally cloned by screening a *C.elegans* N2 genomic DNA lambda library with degenerate oligonucleotides based on the GATA-1 sequence (Spieth *et al.,* 1991). More recently, transformation rescue experiments and sequencing of the *elt-1* coding region from three *C.elegans* strains with non-complementing zygotic lethal mutations which define the *hyd-1* locus, have provided strong evidence that *hyd-1* and *elt-1* are the same gene, and therefore that *elt-1* plays an important role in specification of the hypodermis (B.D.Page, pers. comm.). Furthermore, the *hyd-1* mutation can be phenocopied by transformation of *C.elegans* with constructs generating *elt-1* antisense RNA(K.Steward, pers. comm.). These experiments suggest that *elt-1* plays a crucial role in specification of the hypodermis during embryogenesis. *elt-2* was isolated from *a C.elegans* cDNA expression library using a region of the *ges-1* promoter that possesses a tandem pair of GATA motifs (Hawkins and McGhee, 1995). The *elt-2* gene has subsequently been shown to exhibit gut-specific expression, using *lacZ* reporter constructs (M.Hawkins, pers. comm.). Finally, the *end-1* gene was cloned using a transformation rescue approach. Sequencing of this gene revealed that (like *elt-1*, *elt-2* and *elt-3*) *end-1* encodes a protein with significant homology to the DNA-binding domain of the GATA family of transcription factors. *end-1* is thought to be important for specification of the endoderm, since *end-1* mutants exhibit endodermal defects including a complete absence of the E-lineage (J.Zhu, pers. comm.). Together, these data provide a substantial amount of evidence to suggest that the GATA family of transcription factors are important regulators of a number of lineage restricted genes in *C.elegans*.

It is not possible to determine whether or not putative DNA binding motifs are functional from sequence data alone. However the clustering of a number of classic GATA motifs (and variations of this motif that may be functionally significant) in the 5′ flanks of *cpr-3, cpr-4*, *cpr-5* and *cpr-6* suggest that expression of these genes may also be regulated by the GATA family of transcription factors (Section 4.2.6, Figure 4.17). Furthermore, the presence of a TATA box in the 5′ flank sequences of only *cpr-4* and *cpr-6*, suggests that the mechanism of regulation of these two genes by GATA factors may be distinct from *cpr-3* and *cpr-5*. Indeed the differences in the number, type and position of GATA motifs, as well as the presence or absence of a TATA box, in the 5′ flank regions of *cpr-3, cpr-4*, *cpr-5* and *cpr-6* suggest that these genes may require different combinations of GATA factors for appropriate expression.

It is interesting to note that at least one of the classic GATA motifs present in the 5′ flank of each of the four genes, *cpr-3, cpr-4*, *cpr-5* and *cpr-6*, is a perfect VPE2 sequence. This sequence (CTGATAA) represents a subset of GATA motifs first identified in the promoter regions of all six *vit* genes of *C.elegans* which are expressed exclusively in the intestinal cells of hermaphrodites (Spieth *et al.*, 1985). Subsequently, the *cpr-1* cathepsin B-like gene and the *ges-1* gene, which are both expressed exclusively in the intestinal cells of *C.elegans*, have both been shown to possess at least one copy of the VPE2 motif in their respective promoter regions (Ray and McKerrow, 1992; Egan *et al.,* 1995). Furthermore, mutation analysis of the *vit-2* (MacMorris *et al.,* 1992) and *ges-1* (Egan *et al.,* 1995) promoters have demonstrated that the VPE2 sequences perform important roles in the expression of these two genes. Mutational analysis of the VPE2 sequences of the *vit-2* gene has revealed that these sequences perform an important role in the activation of *vit-2* in the intestine. However, these studies have also demonstrated that the VPE2 sequences may not be absolutely required, since complete deletion of the VPE2 sequences results in very low level, but correctly regulated, expression of the *vit-2* gene (MacMorris *et al.,* 1992). In contrast, the GATA motifs in the promoter region of the *ges-1* gene, including the VPE2 sequence, appear to perform roles as both tissue specific transcriptional enhancers and silencers (Egan *et al.,* 1995) since these sequences are required for activation of *ges-1* expression in the intestine and silencing of *ges-1* expression in the pharynx and tail. Together, these data suggest that VPE2 sequences may perform important roles in the regulation and activation of a number of gut-specific genes in *C.elegans* but that the nature of this regulation may be dependent on the context within which the GATA motifs are found in their respective promoters. Thus, the presence of at least one VPE2 sequence in the promoter regions of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, in conjunction with the known gut-specific expression of *cpr-1*, suggest that the four cathepsin B-like genes reported here may also exhibit gut-specific expression which may in part be regulated by the interaction of different members of the GATA family of transcription factors.

In summary, nucleotide sequence analysis of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, in addition to their mature transcripts, indicates these genes all have very different gene architectures. The predicted protein products of these genes display the features expected of the cathepsin B class of cysteine protease enzyme but may have different substrate specificities and activities from one another. Phylogenetic analysis, and comparison of the predicted protein sequences, indicate that the products of these genes are highly diverged and may therefore have evolved to perform distinct biological functions. Analysis of the 5′ flank sequences of the four genes reveals clustering of several putative GATA regulatory elements. The presence of these elements suggest that these genes may be temporally and/or spatially regulated.

Table 4.1

The table shows the expected sizes of the cloned 5′ RACE products for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as determined from the sizes of the predominant 5′ RACE products visualised by ethidium bromide staining of the 1.5% agarose gel (Figure 4.2). The table also shows the actual size range of the cloned 5′ RACE products for each gene, determined from their nucleotide sequences.

Table 4.1

Gene	Expected size of RACE products, determined from agarose gel electrophoresis	Size of 5' RACE products, determined from DNA sequence analysis
cpr-3	250 bp	225 bp - 278 bp
cpr-4	210 bp	207 bp -. 213 bp
cpr-5	330 bp	328 bp - 343 bp
cpr-6	270 bp and 330 bp	274 bp - 287 bp

Table 4.2

The table summarises the data obtained using the COMPARE and DOTPLOT programs of the GCG package. These programs were used to compare the nucleotide sequences of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to one another. The regions of homology between genes identified using the two programs are given in the table. For each region of homology, the approximate nucleotide positions delineating the beginning and end of the region are given. The numbers indicate the nucleotide position with respect to the putative ATG initiation codon for the appropriate gene (set at +1).

Table 4.2

gene	homology region (bp)	gene	homology region (bp)
cpr-4	+240 - +490 +490 - +790 +790 - +1040	*cpr-5*	+230 - +480 +580 - +830 +930 - +1280
cpr-1	+170 - +250 +300 - +550 +800 - +1100	*cpr-4*	+120 - +190 +240 - +540 +690 - +1040
cpr-1	+150 - +300 +300 - +600 +700 - +1100	*cpr-3*	+130 - +280 +330 - +680 +1180 - +1580
cpr-1	+150 - +300 +300 - +550 +800 - +1100	*cpr-5*	+80 - +230 +230 - +480 +980 - +1280
cpr-4	+140 - +200 +240 - +540 +790 - +1040	*cpr-6*	+230 - +300 +630 - +980 +1180 - +1580
cpr-1	+300 - +650 +850 - +1120	*cpr-6*	+580 - +930 +1180 - +1530
cpr-3	+280 - +580 +1130 - +1480	*cpr-4*	+240 - +540 +740 - +1040
cpr-3	+230 - +580 +1180 - +1480	*cpr-5*	+180 - +480 +980 - +1280
cpr-6	+230 - +330 +630 - +880 +1180 - +1530	*cpr-5*	+80 - +160 +230 - +480 +980 - +1280
cpr-6	+630 - +930 +1280 - +1530	*cpr-3*	+330 - +630 +1230 - +1480

Table 4.3

The table summarises the data obtained using the GAP program of the GCG package. The program was used to generate DNA sequence alignments of the different regions of homology between *cpr*-1, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, identified using the COMPARE and DOTPLOT programs (Table 4.2). The table shows both the maximum percentage homology and the maximum length of uninterrupted homology observed between each region analysed.

Table 4.3

gene	region of gene	vs.	gene	region of gene	region of maximum homology	longest length of continuous homology
cpr-3	+130 - +280		*cpr-1*	+150 - +300	83.3% (25/30)	8 bp
	+330 - +680			+300 - +600	91.5% (43/47)	20 bp
	+1180 - +1580			+700 - +1100	88.9% (16/18)	11 bp
cpr-6	+230 - +300		*cpr-4*	+140 - +200	81.8% (9/11)	4 bp
	+630 - +980			+240 - +540	95.7% (22/23)	18 bp
	+1180 - +1580			+790 - +1040	87.9% (29/33)	11 bp
cpr-1	+150 - +300		*cpr-5*	+80 - +230	85.7% (18/21)	11 bp
	+300 - +550			+230 - +480	89.5% (17/19)	12 bp
	+800 - +1100			+980 - +1280	91.9% (34/37)	16 bp
cpr-1	+170 - +250		*cpr-4*	+120 - +190	86.7% (13/15)	6 bp
	+300 - +550			+240 - +540	87.5% (35/40)	18 bp
	+800 - +1100			+690 - +1040	100% (29/29)	29 bp
cpr-5	+230 - +480		*cpr-4*	+240 - +490	93.8% (30/32)	17 bp
	+580 - +830			+490 - +790	93.3% (42/45)	23 bp
	+930 - +1280			+790 - +1040	96.8% (30/31)	25 bp
cpr-3	+280 - +580		*cpr-4*	+240 - +540	94.7% (18/19)	14 bp
	+1130 - +1480			+740 - +1040	80.8% (21/26)	11 bp
cpr-3	+230 - +580		*cpr-5*	+180 - +480	91.3% (21/23)	11 bp
	+1180 - +1480			+980 - +1280	87.0% (20/23)	11 bp
cpr-6	+230 - +330		*cpr-5*	+80 - +160	100% (17/17)	17 bp
	+630 - +880			+230 - +480	93.8% (15/16)	11 bp
	+1180 - +1530			+980 - +1280	90.6% (29/32)	18 bp
cpr-6	+580 - +930		*cpr-1*	+300 - +650	100% (27/27)	27 bp
	+1180 - +1530			+850 - +1120	87.5% (28/32)	12 bp
cpr-6	+630 - +930		*cpr-3*	+330 - +630	92.6% (25/27)	20 bp
	+1280 - +1530			+1230 - +1480	84.2% (16/19)	8 bp

Table 4.4

The table shows the percentage similarity (SIM) and percentage identity (ID) between the prepro- enzyme and predicted mature enzyme forms of the cathepsin B-like gene products of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* and of human cathepsin B. The data were obtained from alignments of the amino acid sequence of each enzyme pair using the GAP program of the GCG package.

Table 4.4

		cpr-4		*cpr-5*		*cpr-6*		*cpr-1*		human cathepsin B	
		SIM	ID	SI M	ID	SI M	ID	SIM	ID	SIM	ID
cpr-3	Mature	**68.9**	**54.5**	**69.0**	**56.9**	**68.6**	**51.0**	**75.0**	**66.0**	**68.1**	**54.2**
	Pre-Pro	**64.2**	**47.8**	**65.2**	**50.9**	**63.6**	**43.4**	**68.4**	**57.4**	**63.9**	**45.8**
cpr-4	Mature			**83.5**	**77.3**	**69.8**	**56.1**	**72.5**	**61.9**	**72.2**	**56.3**
	Pre-Pro			**76.9**	**67.6**	**62.8**	**47.6**	**70.1**	**56.2**	**64.1**	**47.7**
cpr-5	Mature					**71.0**	**57.5**	**70.6**	**62.0**	**69.3**	**55.6**
	Pre-Pro					**67.1**	**52.9**	**66.6**	**58.6**	**63.2**	**48.8**
cpr-6	Mature							**70.5**	**56.6**	**70.7**	**53.7**
	Pre-Pro							**63.7**	**47.6**	**66.8**	**47.6**
cpr-1	Mature									**69.8**	**57.8**
	Pre-Pro									**62.3**	**47.8**

Figure 4.1

A). Summary of the 5′ RACE methodology

1 Anneal pooled gene specific primers (GSP1) for each of the four genes to total RNA from a mixed stage population of *C.elegans*.

2 Synthesise 1st strand cDNA using Superscript II reverse transcriptase.

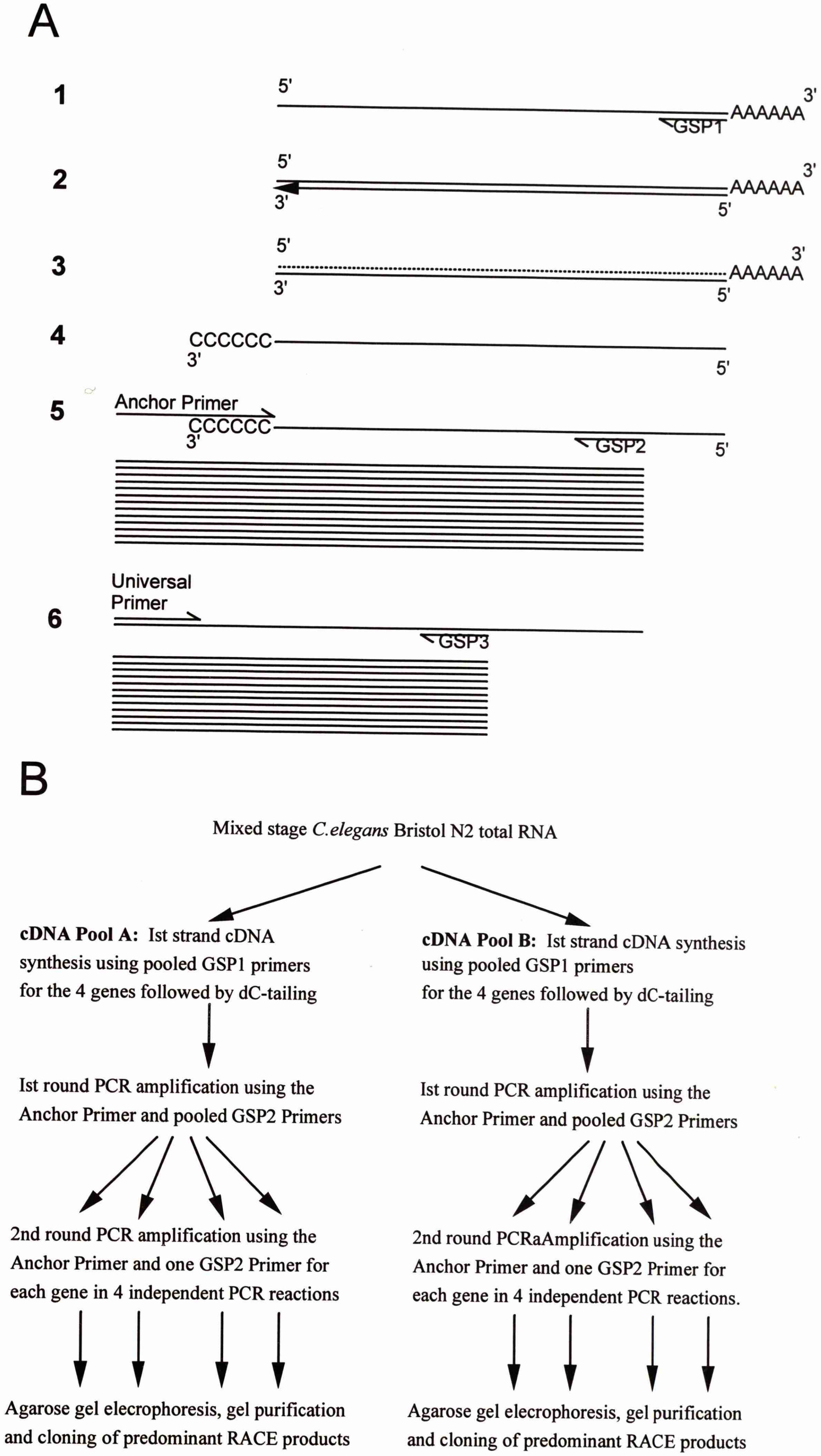
3 Degrade RNA using RNase H.

4 Generate dC-tailed 1st strand cDNA using dCTP and terminal deoxynucleotide transferase.

5 **1st Round PCR:** PCR amplify the dC-tailed 1st strand cDNA using pooled gene specific primers (GSP2) for each of the four genes and the Anchor primer.

6 **2nd Round PCR:** PCR amplify the products from the 1st Round PCR reaction using the Universal primer and one gene specific primer (GSP3) in four independent PCR reactions (1 reaction for each gene).

B). A summary of the 5′ RACE protocol



Mixed stage *C.elegans* Bristol N2 total RNA

cDNA Pool A: Ist strand cDNA synthesis using pooled GSP1 primers for the 4 genes followed by dC-tailing

Ist round PCR amplification using the Anchor Primer and pooled GSP2 Primers

2nd round PCR amplification using the Anchor Primer and one GSP2 Primer for each gene in 4 independent PCR reactions

Agarose gel elecrophoresis, gel purification and cloning of predominant RACE products

cDNA Pool B: Ist strand cDNA synthesis using pooled GSP1 primers for the 4 genes followed by dC-tailing

Ist round PCR amplification using the Anchor Primer and pooled GSP2 Primers

2nd round PCRaAmplification using the Anchor Primer and one GSP2 Primer for each gene in 4 independent PCR reactions.

Agarose gel elecrophoresis, gel purification and cloning of predominant RACE products

Figure 4.2

Results obtained from the 5′ RACE experiments

A. Results obtained from PCR amplification of cDNA pool A (see Figure 4.1B).

B. Results obtained from PCR amplification of cDNA pool B (see Figure 4.1B).

Lane 1: PCR products generated by the 1st round PCR amplification step

Lanes 2-5: PCR products obtained after the 2nd Round PCR amplification step

Lane 2:	*cpr-3*
Lane 3:	*cpr-6*
Lane 4:	*cpr-4*
Lane 5:	*cpr-5*

C. GibcoBRL 5′ RACE controls

Lane X: Results obtained after the 1st round PCR amplification step using the Anchor Primer, the GibcoBRL control GSP2 primer and dC-tailed cDNA template generated from 1 pg of GibcoBRL control RNA (an *in-vitro* transcribed 891bp region of the Chloramphenicol Acetyl Transerase gene engineered to include a poly(A) Tail). A 2nd round PCR amplification step was not performed with the GibcoBRL control.

Expected size of product from GibcoBRL control RNA: 738bp

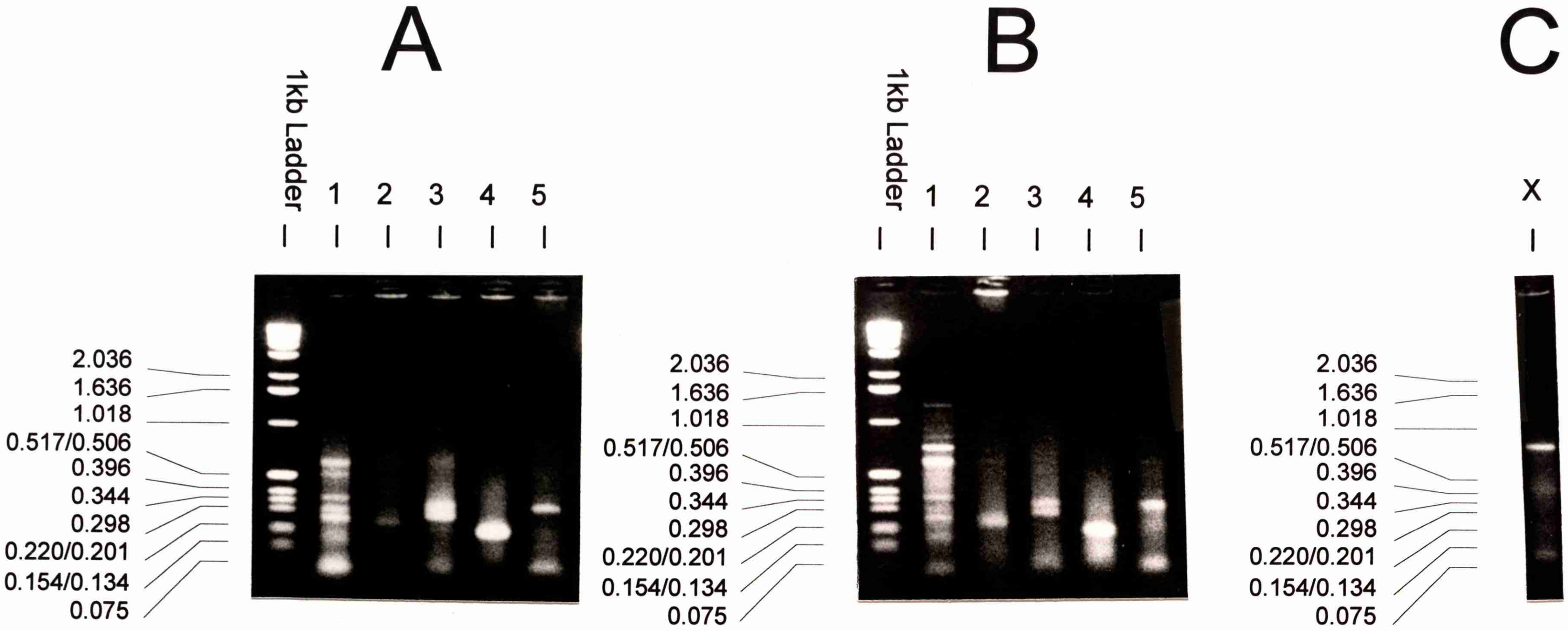

A
B
C
1kb Ladder
1
2
3
4
5
1kb Ladder
1
2
3
4
5
x
2.036
1.636
1.018
0.517/0.506
0.396
0.344
0.298
0.220/0.201
0.154/0.134
0.075
2.036
1.636
1.018
0.517/0.506
0.396
0.344
0.298
0.220/0.201
0.154/0.134
0.075
2.036
1.636
1.018
0.517/0.506
0.396
0.344
0.298
0.220/0.201
0.154/0.134
0.075

Figure 4.3

The results obtained after repeating the 2nd round PCR amplification step for *cpr-6*

The 2nd round PCR amplification step was performed using the 1st round PCR amplification products from cDNA pool A (see Figure 4.1 and 4.2), the Universal Primer and the *cpr-6* GSP3 primer.

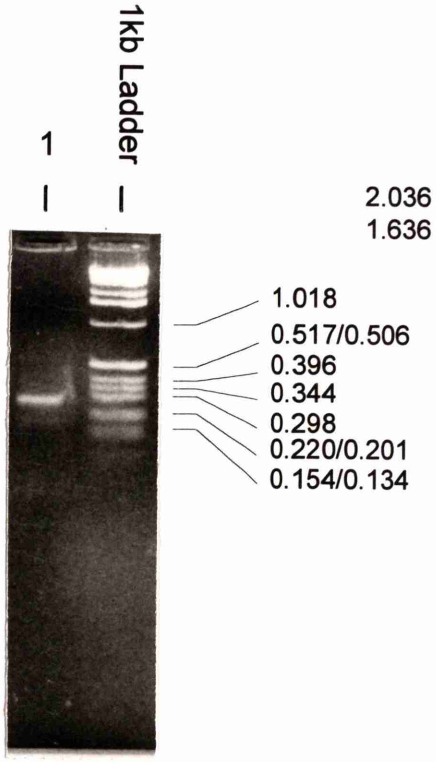
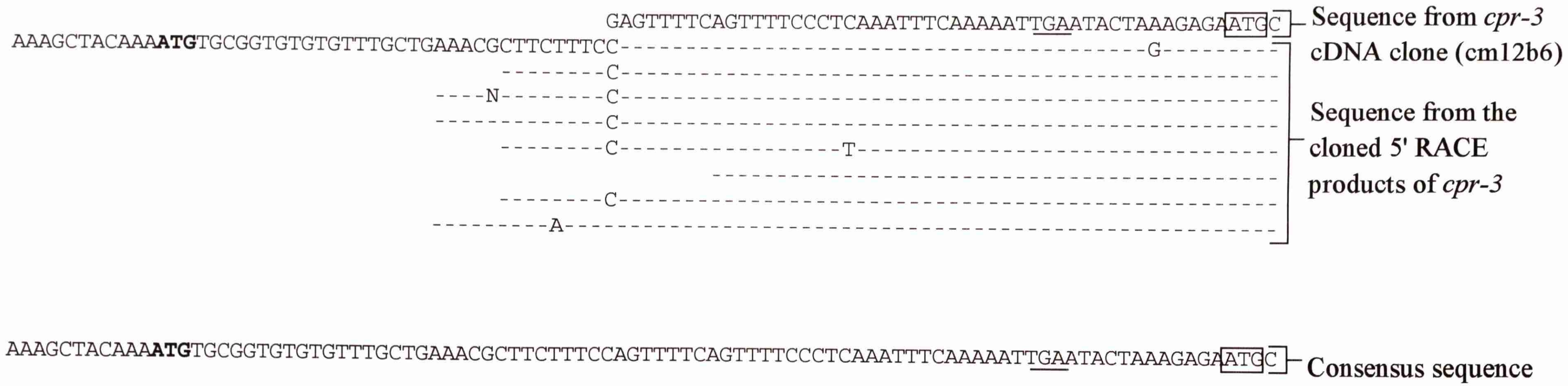

1
1kb Ladder
2.036
1.636
1.018
0.517/0.506
0.396
0.344
0.298
0.220/0.201
0.154/0.134

Figure 4.4

DNA Sequence of the cloned 5′ RACE products of *cpr-3*

The figure shows the DNA sequence of each of the 5′ RACE clones of *cpr-3* analysed, including a 50bp region of overlap with the *cpr-3* cDNA clone (cm12b6). The nucleotide sequence of the 50bp region of the cm12b6 cDNA clone to which the 5′ RACE clones were compared and the nucleotide sequence of the largest 5′ RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5′ RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG). A second ATG codon, in-frame with the putative ATG initiation codon, which was identified in one of the eight sequenced 5′ RACE clones is marked in bold (**ATG**). A translation termination codon in-frame with and upstream of the putative ATG translation initiation codon is underlined (TGA)



```
                                             GAGTTTTCAGTTTTCCCTCAAATTTCAAAAATTGAATACTAAAGAGAATGC
AAAGCTACAAAATGTGCGGTGTGTGTTTGCTGAAACGCTTCTTTCC-----------------------------------------G--------
                                     --------C--------------------------------------------------
                                ----N--------C--------------------------------------------------
                                -------------C--------------------------------------------------
                                     --------C-----------------T--------------------------------
                                                     -------------------------------------------
                                     --------C--------------------------------------------------
                                ---------A------------------------------------------------------

AAAGCTACAAAATGTGCGGTGTGTGTTTGCTGAAACGCTTCTTTCCAGTTTTCAGTTTTCCCTCAAATTTCAAAAATTGAATACTAAAGAGAATGC
```

Sequence from *cpr-3* cDNA clone (cm12b6)
Sequence from the cloned 5' RACE products of *cpr-3*
Consensus sequence

Figure 4.5

DNA Sequence of the cloned 5′ RACE products of *cpr-4*

The figure shows the DNA sequence of each of the 5′ RACE clones of *cpr-4* analysed, including a 50bp region of overlap with the *cpr-4* cDNA clone (cm14e3). The nucleotide sequence of the 50bp region of the cm14e3 cDNA clone to which the 5′ RACE clones were compared and the nucleotide sequence of the largest 5′ RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5′ RACE clone sequences. The putative ATG translation initiation codon is boxed (ATG)

```
                 GGCTATTTGCTCTTTTACAAAAATGAAATACCTCATTCTTGCTGCCTTGGT  Sequence from cpr-4
CAGCATTTGCTCTATCTT--------------------------------------------------  cDNA clone (cm14e3)
   --------------T-----G---------------G----------------------------
   --------------T--------------------------------------------------  Sequence from the
   --------------T--------------------------------------------------  cloned 5' RACE
      -----------T--------------------------------------------------  products of cpr-4
   -T------------T--------------------------------------------------


CAGCATTTGCTCTATCTTGCTATTTGCTCTTTTACAAAAATGAAATACCTCATTCTTGCTGCCTTGGT  Consensus sequence
```

Figure 4.6

DNA Sequence of the cloned 5′ RACE products of *cpr-5*

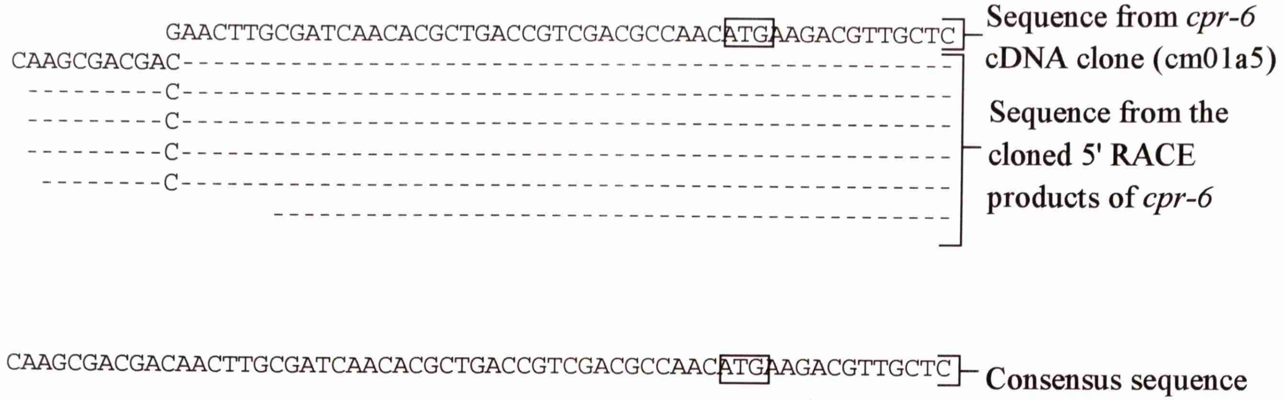
The figure shows the DNA sequence of each of the 5′ RACE clones of *cpr-5* analysed, including a 50bp region of overlap with the *cpr-5* cDNA clone (cm04d10). The nucleotide sequence of the 50bp region of the cm04d10 cDNA clone to which the 5′ RACE clones were compared and the nucleotide sequence of the largest 5′ RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the appropriate region of the 5′ RACE clone sequences. The putative ATG translation initiation codon is boxed ($\boxed{\texttt{ATG}}$).

```
                                                                                                    GCC  Sequence from cpr-5
CGCGAACCTTCGCAGACACTTCTCTCATAATGTGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCTGCCTCGGCTGTGGTGATTCCTGGACACCGTGAAGCC--  cDNA clone (cm04d10)
        ---------T--------------A-N-----------------------------------------------------------------C--
             ---------A-C---------------------------------------------------------------------------C--  Sequence from the
            ----------------------------------------------------------------------------------------C--  cloned 5' RACE
               -------------------------------------------------------------N-----------------------C--  products of cpr-5

CGCGAACCTTCGCAGACACTTCTCTCATAATGTGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCTGCCTCGGCTGTGGTGATTCCTGGACACCGTGAAGCCCC  Consensus sequence



                                                      AGCTCTCACTGGACAAGCTTTGATCGACTATGTCAACTCTGCCCAAAA  Sequence from cpr-5
                                                      ------------------------------------------------  cDNA clone (cm04d10)
                                                      ------------------------------------------------
                                                      ------------------------------------------------  Sequence from the
                                                      ------------------------------------------------  cloned 5' RACE
                                                      ------------------------------------------------  products of cpr-5

                                                      AGCTCTCACTGGACAAGCTTTGATCGACTATGTCAACTCTGCCCAAAA  Consensus sequence
```

Figure 4.7

DNA Sequence of the cloned 5′ RACE products of *cpr-6*

The figure shows the DNA sequence of each of the 5′ RACE clones of *cpr-6* analysed, including a 50bp region of overlap with the *cpr-6* cDNA clone (cm01a5). The nucleotide sequence of the 50bp region of the cm01a5 cDNA clone to which the 5′ RACE clones were compared and the nucleotide sequence of the largest 5′ RACE clone sequenced is given. Thereafter, only sequence mismatches are shown, with identical bases being marked by a dash (-). The consensus sequence generated from these data is shown below the 5′ RACE clone sequences. The putative ATG translation initiation codon is boxed ($\boxed{\texttt{ATG}}$).



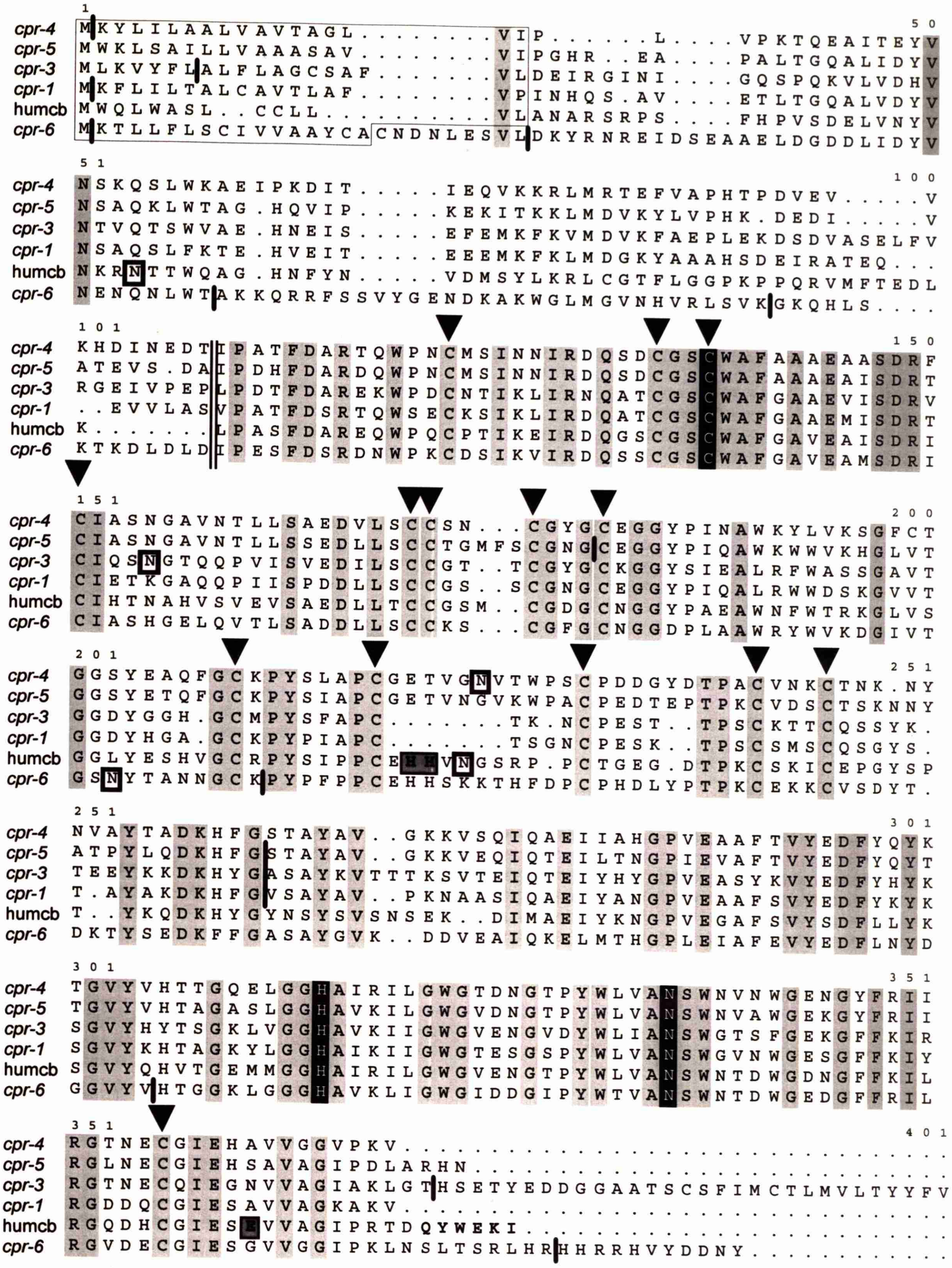
```
          GAACTTGCGATCAACACGCTGACCGTCGACGCCAACATGAAGACGTTGCTC
CAAGCGACGAC--------------------------------------------------
----------C--------------------------------------------------
----------C--------------------------------------------------
----------C--------------------------------------------------
  --------C--------------------------------------------------
                 --------------------------------------------

CAAGCGACGACAACTTGCGATCAACACGCTGACCGTCGACGCCAACATGAAGACGTTGCTC
```

Sequence from *cpr-6* cDNA clone (cm01a5)
Sequence from the cloned 5' RACE products of *cpr-6*
Consensus sequence

Figure 4.8

A comparison of the predicted amino acid sequences of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, together with the previously isolated *cpr-1* cathepsin B-like gene and human cathepsin B (humcb)

The alignment was generated using the PILEUP program of the GCG package and some alignment by eye. The region boxed at the N-terminus indicates the hydrophobic residues of the putative pre- domain.



```
      1                                                50
cpr-4 MKYLILAALVAVTAGL........VIP......L....VPKTQEAITEYV
cpr-5 MWKLSAILLVAAASAV........VIPGHR..EA....PALTGQALIDYV
cpr-3 MLKVYFLALFLAGCSAF.......VLDEIRGINI....GQSPQKVLVDHV
cpr-1 MKFLILTALCAVTLAF........VPINHQS.AV....ETLTGQALVDYV
humcb MWQLWASL..CCLL..........VLANARSRPS....FHPVSDELVNYV
cpr-6 MKTLLFLSCIVVAAYCACNDNLESVLDKYRNREIDSEAAELDGDDLIDYV

      51                                              100
cpr-4 NSKQSLWKAEIPKDIT.....IEQVKKRLMRTEFVAPHTPDVEV.....V
cpr-5 NSAQKLWTAG.HQVIP.....KEKITKKLMDVKYLVPHK.DEDI.....V
cpr-3 NTVQTSWVAE.HNEIS.....EFEMKFKVMDVKFAEPLEKDSDVASELFV
cpr-1 NSAQSLFKTE.HVEIT.....EEEMKFKLMDGKYAAAHSDEIRATEQ...
humcb NKRNTTWQAG.HNFYN.....VDMSYLKRLCGTFLGGPKPPQRVMFTEDL
cpr-6 NENQNLWTAKKQRRFSSVYGENDKAKWGLMGVNHVRLSVKGKQHLS....

      101                                             150
cpr-4 KHDINEDTIPATFDARTQWPNCMSINNIRDQSDCGSCWAFAAAEAASDRF
cpr-5 ATEVS.DAIPDHFDARDQWPNCMSINNIRDQSDCGSCWAFAAAEAISDRT
cpr-3 RGEIVPEPLPDTFDAREKWPDCNTIKLIRNQATCGSCWAFGAAEVISDRV
cpr-1 ..EVVLASVPATFDSRTQWSECKSIKLIRDQATCGSCWAFGAAEMISDRT
humcb K.......LPASFDAREQWPQCPTIKEIRDQGSCGSCWAFGAVEAISDRI
cpr-6 KTKDLDLDIPESFDSRDNWPKCDSIKVIRDQSSCGSCWAFGAVEAMSDRI

      151                                             200
cpr-4 CIASNGAVNTLLSAEDVLSCCSN...CGYGCEGGYPINAWKYLVKSGFCT
cpr-5 CIASNGAVNTLLSSEDLLSCCTGMFSCGNGCEGGYPIQAWKWWVKHGLVT
cpr-3 CIQSNGTQQPVISVEDILSCCGT..TCGYGCKGGYSIEALRFWASSGAVT
cpr-1 CIETKGAQQPIISPDDLLSCCGS..SCGNGCEGGYPIQALRWWDSKGVVT
humcb CIHTNAHVSVEVSAEDLLTCCGSM..CGDGCNGGYPAEAWNFWTRKGLVS
cpr-6 CIASHGELQVTLSADDLLSCCKS...CGFGCNGGDPLAAWRYWVKDGIVT

      201                                             251
cpr-4 GGSYEAQFGCKPYSLAPCGETVGNVTWPSCPDDGYDTPACVNKCTNK.NY
cpr-5 GGSYETQFGCKPYSIAPCGETVNGVKWPACPEDTEPTPKCVDSCTSKNNY
cpr-3 GGDYGGH.GCMPYSFAPC.......TK.NCPEST..TPSCKTTCQSSYK.
cpr-1 GGDYHGA.GCKPYPIAPC.......TSGNCPESK..TPSCSMSCQSGYS.
humcb GGLYESHVGCRPYSIPPCEHHVNGSRP.PCTGEG.DTPKCSKICEPGYSP
cpr-6 GSNYTANNGCKPYPFPPCEHHSKKTHFDPCPHDLYPTPKCEKKCVSDYT.

      251                                             301
cpr-4 NVAYTADKHFGSTAYAV..GKKVSQIQAEIIAHGPVEAAFTVYEDFYQYK
cpr-5 ATPYLQDKHFGSTAYAV..GKKVEQIQTEILTNGPIEVAFTVYEDFYQYT
cpr-3 TEEYKKDKHYGASAYKVTTTKSVTEIQTEIYHYGPVEASYKVYEDFYHYK
cpr-1 T.AYAKDKHFGVSAYAV..PKNAASIQAEIYANGPVEAAFSVYEDFYKYK
humcb T..YKQDKHYGYNSYSVSNSEK..DIMAEIYKNGPVEGAFSVYSDFLLYK
cpr-6 DKTYSEDKFFGASAYGVK..DDVEAIQKELMTHGPLEIAFEVYEDFLNYD

      301                                             351
cpr-4 TGVYVHTTGQELGGHAIRILGWGTDNGTPYWLVANSWNVNWGENGYFRII
cpr-5 TGVYVHTAGASLGGHAVKILGWGVDNGTPYWLVANSWNVAWGEKGYFRII
cpr-3 SGVYHYTSGKLVGGHAVKIIGWGVENGVDYWLIANSWGTSFGEKGFFKIR
cpr-1 SGVYKHTAGKYLGGHAIKIIGWGTESGSPYWLVANSWGVNWGESGFFKIY
humcb SGVYQHVTGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKIL
cpr-6 GGVYVHTGGKLGGGHAVKLIGWGIDDGIPYWTVANSWNTDWGEDGFFRIL

      351                                              401
cpr-4 RGTNECGIEHAVVGGVPKV................................
cpr-5 RGLNECGIEHSAVAGIPDLARHN............................
cpr-3 RGTNECQIEGNVVAGIAKLGTHSETYEDDGGAATSCSFIMCTLMVLTYYFV
cpr-1 RGDDQCGIESAVVAGKAKV................................
humcb RGQDHCGIESEVVAGIPRTDQYWEKI.........................
cpr-6 RGVDECGIESGVVGGIPKLNSLTSRLHRHHRRHVYDDNY............
```

KEY: | , Intron position ; X , Conserved residue; N , Potential N-glycosylation site; ▼ , Conserved cysteine residue; X , Catalytic residue; || , Potential proregion cleavage site; X, Residue which may affect enzyme activity/specificity

Figure 4.9

A comparison of the 5′ flank sequence of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to the consensus DNA sequence generated by sequencing the 5′ RACE clones for each of the four genes

The DNA sequences of the 5′ RACE clones were compared to the DNA sequences of the 5' flank of the appropriate gene, up to the putative ATG translation initiation codon.

```
cpr-3     GATAAGGGAAAGCTACAAAATGTGCGGTGTGTGTTTGCTGAAACGCTTCTTTCCAGTTTTCAGTTTTCCCTCAAATTTCAAAAATTGAATACTAAAGAGA ATG
                  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
                  AAAGCTACAAAATGTGCGGTGTGTGTTTGCTGAAACGCTTCTTTCCAGTTTTCAGTTTTCCCTCAAATTTCAAAAATTGAATACTAAAGAGA ATG


cpr-4     AACGAGCACTCCCGAACTGATAACGAGTCAACTATAAAAGACCATCGCAATGAAGTAACTTCAGCATTTGCTCTATCTTGCTATTTGCTCTTTTACAAAA ATG
                                                                       ||||||||||||||||||||||||||||||||||||||| |||
                                                                       CAGCATTTGCTCTATCTTGCTATTTGCTCTTTTACAAAA ATG


cpr-5     ATTCTTAAAAAAGATACTGATAAAATTAACTGATAAGAATGATGCGTACCGACTACTTAATGAGCATTAGACGCGAACCTTCGCAGACACTTCTCTCATA ATG
                                                                                 ||||||||||||||||||||||||||||| |||
                                                                                 CGCGAACCTTCGCAGACACTTCTCTCATA ATG


cpr-6     AAATTTGATAACGAAAACATGTTCTATAAAAGCATGCTGATAAAAGCGAGCAGTCAAGCGACGACAACTTGCGATCAACACGCTGACCGTCGACGCCAAC ATG
                                                                ||||||||||||||||||||||||||||||||||||||||||||||| |||
                                                                CAAGCGACGACAACTTGCGATCAACACGCTGACCGTCGACGCCAAC ATG
                                                                                                               ↑
```

Translation Initiation Codon

For each gene; **the upper sequence represents the 5' flank sequence of the genomic clones**
the lower sequence represents the consensus sequence generated from the cloned 5' RACE products

Figure 4.10

Potential RNA secondary structure for the 3′ UTR of the mature transcript of *cpr-3*

The structure was calculated using the FOLDRNA program of the GCG package. The free energy for the secondary structure was calculated in kilocalories (kcal) at 37°C by the FOLDRNA program.

Energy: -20.1

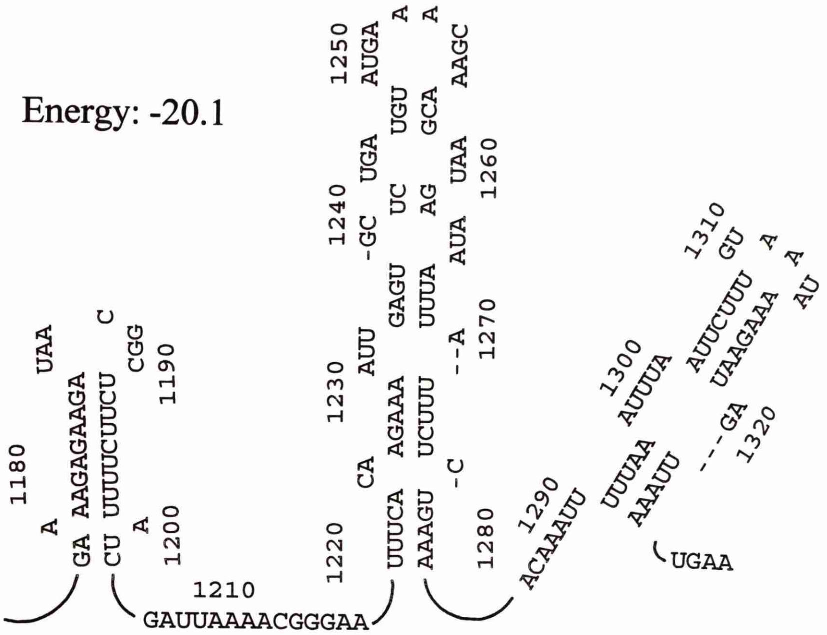

1180
A
GA AAGAGAAGA
UAA
CU UUUUCUUCU
C
A
CGG
1190
1200
1210
GAUUAAAACGGGAA
1220
UUUCA
AAAGU
CA
AGAAA
UCUUU
-C
1230
AUU
GAGU
UUUA
--A
1270
1240
-GC
UC
AG
AUA
UGA
UAA
1260
UGU
GCA
1250
AUGA
AAGC
A
A
1280
1290
ACAAAUU
UUUAA
AAAUU
1300
AUUUA
AUUCUUU
UAAGAAA
---GA
1320
1310
GU
A
A
AU
UGAA

Figure 4.11

The annotated DNA sequence of *cpr-3*

The coding DNA sequence is indicated in capitals (CAPITALS). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (▒). Non-coding sequences (introns and 5′ and 3′ flank sequences) are indicated in lowercase (lowercase). The 5′ end of each sequenced 5′ RACE clone is indicated by a vertical arrow (↓). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. The double vertical lines (‖) indicate the start of the poly(A) tail and the boxed nucleotides (▭) indicate the putative cleavage/poly(A) signal.

```
                                                     1
                                                     ↓
-140 aatagacaaacataagaggggaaaatgtcgaaatttagtgataagggaaagctacaaaa -81
```

```
-320 tccagacaatttaccctaaagtgtgaccaaatgcctacattacacacacctcaacgcttt -261
-260 gataagcctgataagctccctcctaattcataatgatgatatcaacgaaggtgataattg -201
-200 atttcttgattcggtgacactaacctccgtgttctcgactgatgactaactttttttttg -141
                                                          1
                                                          ↓
-140 aatagacaaacataagaggggaaaatgtcgaaatttagtgataagggaaagctacaaaa -81
                         3    3               1
                         ↓    ↓               ↓
 -80 tgtgcggtgtgtgtttgctgaaacgcttctttccagttttcagttttccctcaaatttca -21
 -20 aaaattgaatactaaagagaATGCTGAAAGTGTACTTTTTGgtgagtcatcttcatcttc 40
                          M  L  K  V  Y  F  L
  41 atttttccattatttttcttacatttagGCACTGTTTCTAGCCGGGTGCTCTGCATTTG 100
                                  A  L  F  L  A  G  C  S  A  F  V
 101 TTCTTGATGAAATCCGCGGAATCAATATTGGGCAGTCACCTCAGAAAGTCCTTGTAGATC 160
       L  D  E  I  R  G  I  N  I  G  Q  S  P  Q  K  V  L  V  D  H
 161 ATGTGAACACCGTACAAACCAGCTGGGTGGCAGAGCACAATGAGATTTCCGAGTTTGAGA 220
       V  N  T  V  Q  T  S  W  V  A  E  H  N  E  I  S  E  F  E  M
 221 TGAAGTTCAAAGTGATGGATGTGAAGTTCGCAGAGCCTTTGGAAAAAGATTCCGATGTGG 280
       K  F  K  V  M  D  V  K  F  A  E  P  L  E  K  D  S  D  V  A
 281 CCAGTGAGCTTTTCGTTAGAGGAGAGATTGTTCCAGAACCACTCCCCGACACTTTTGACG 340
       S  E  L  F  V  R  G  E  I  V  P  E  P  L  P  D  T  F  D  A
 341 CCAGAGAAAAATGGCCAGACTGTAATACAATAAAGCTTATCCGAAACCAGGCCACCTGCG 400
       R  E  K  W  P  D  C  N  T  I  K  L  I  R  N  Q  A  T  C  G
 401 GATCCTGCTGGGCTTTCGGTGCGGCAGAGGTGATTTCCGACCGAGTGTGCATTCAGTCTA 460
       S  C  W  A  F  G  A  A  E  V  I  S  D  R  V  C  I  Q  S  N
 461 ATGGAACTCAACAACCTGTGATCTCAGTTGAAGATATTCTTTCTTGTTGTGGAACAACCT 520
       G  T  Q  Q  P  V  I  S  V  E  D  I  L  S  C  C  G  T  T  C
 521 GTGGTTACGGATGTAAGGGAGGCTACTCAATTGAGGCGTTGCGTTTCTGGGCCAGTAGTG 580
       G  Y  G  C  K  G  G  Y  S  I  E  A  L  R  F  W  A  S  S  G
 581 GAGCAGTAACAGGTGGAGATTACGGAGGACACGGATGTATGCCTTACTCGTTTGCTCCGT 640
       A  V  T  G  G  D  Y  G  G  H  G  C  M  P  Y  S  F  A  P  C
 641 GCACGAAGAATTGCCCGGAGTCAACGACTCCAAGCTGTAAGACAACTTGCCAATCTAGCT 700
       T  K  N  C  P  E  S  T  T  P  S  C  K  T  T  C  Q  S  S  Y
 701 ACAAAACGGAAGAATACAAAAAGGATAAGCATTATGgtgagttggtctggcattcattta 760
       K  T  E  E  Y  K  K  D  K  H  Y  G
 761 atagatttcagaggtttctgaacagtttttttaaattttgtgtttacgagaatttttatc 820
 821 agatgaaataatgtatttcaaaaatttagttgccctcttttagaataatattttttttctg 880
 881 atcgggggcagatagctcagtcggtagtggtggccgctagcagtctggaggtcacgagtt 940
 941 caagtccggcctcacccccctaggttcacccagcctctattgggaagtggagcaatacacg 1000
1001 actggattatcggccacagtccccggctaggacgtggcttaaattacagcgtacctgaat 1060
1061 cccagatccgcagtgcatagcacttgaagaacggatcgtcctttaatcttttaatccttt 1120
1121 aactaacattgtttgtttccagGAGCTTCTGCCTACAAAGTCACCACTACGAAATCCGTC 1180
                             A  S  A  Y  K  V  T  T  T  K  S  V
1181 ACCGAAATTCAAACTGAAATCTACCATTACGGACCTGTGGAGGCTTCATACAAAGTCTAC 1240
     T  E  I  Q  T  E  I  Y  H  Y  G  P  V  E  A  S  Y  K  V  Y
1241 GAGGATTTTTATCATTATAAATCAGGAGTTTATCACTATACTTCCGGAAAGCTTGTTGGA 1300
     E  D  F  Y  H  Y  K  S  G  V  Y  H  Y  T  S  G  K  L  V  G
1301 GGTCACGCGGTTAAAATTATCGGCTGGGGAGTTGAAAACGGAGTGGACTACTGGCTGATT 1360
     G  H  A  V  K  I  I  G  W  G  V  E  N  G  V  D  Y  W  L  I
1361 GCAAATTCCTGGGGCACCAGTTTCGGCGAAAAGGGGTTCTTCAAAATTCGCCGTGGAACA 1420
     A  N  S  W  G  T  S  F  G  E  K  G  F  F  K  I  R  R  G  T
1421 AATGAATGTCAAATTGAAGGAAACGTGGTAGCTGGAATTGCAAAATTGGGAACGgtacgt 1480
     N  E  C  Q  I  E  G  N  V  V  A  G  I  A  K  L  G  T
1481 gaactttgaaaggactttttaaaaatctaaaatttcagCACTCGGAAACGTATGAAGACG 1540
                                           H  S  E  T  Y  E  D  D
1541 ATGGCGGTGCAGCGACCTCGTGTAGTTTTATTATGTGTACATTGATGGTGCTGACGTATT 1600
       G  G  A  A  T  S  C  S  F  I  M  C  T  L  M  V  L  T  Y  Y
1601 ATTTCGTGTAGaaaagagaagataacggctcttcttttatcgattaaaacgggaatttca 1660
       F  V  *
1661 caagaaaattgagtgctctgatgtatgaaacgaaacgaatgaataatttatttctctgaa 1720
1721 aacaaattttttaaatttaattctttgtaataaaagaatagttaaatgaagcaaaacacat 1780
1781 ttattcgagtacaattctaaatagacctggctacgctacatatccggcggagctgcagaa 1840
1841 ccacatgcacttcccggtggacacggttctgggactttttcaattgcatcagcaaacaga 1900
1901 ggaagaagaaccaacatgacgagcaatgcccatttaacaacttctttcatcttgggatgc 1960
1961 agggtggatttgttttgaaaacaaattgactgatttttactattgatcaaatggttatat 2020
2021 atcagagaagttcattgaa 2039
```

Figure 4.12

The annotated DNA sequence of *cpr-4*

The coding DNA sequence is indicated in capitals (CAPITALS). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (▒). Non-coding sequences (introns and 5′ and 3′ flank sequences) are indicated in lowercase (`lowercase`). The 5′ end of each sequenced 5′ RACE clone is indicated by a vertical arrow (↓). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. The double vertical lines (‖) indicate the start of the poly(A) tail and the boxed nucleotides (□) indicate the putative cleavage/poly(A) signal.

```
-359 tagtcctcatagttcaaaccttcttgctacttttacacctaacctaaaaattgagtactt -300
-299 ctaatctggttccaaatgataactttcgttgaaccacacaacttccaaactcttatcaaa -240
-239 gttgcacgagatcattgtgctcaaatgatggtgctgcgtcacatgactacctcctaatta -180
-179 ggcattgtctatcgaaatttgcgctgccaggtaccgcaaattttttcaattttaatcccg -120
-119 gacgccaatgataagagataacgagcactcccgaactgataacgagtcaactataaaaga -60
                         1  4  1
                         ↓  ↓  ↓
 -59 ccatcgcaatgaagtaacttcagcatttgctctatcttgctatttgctcttttacaaaaA 1
                                                                M
   2 TGgtaagtggttgttttagcgtgaagtttaataattgtaattgtcttgaagAAATACCTC 61
                                                         K  Y  L
  62 ATTCTTGCTGCCTTGGTGGCTGTTACCGCCGGACTCGTTATTCCACTTGTTCCAAAAACC 121
      I  L  A  A  L  V  A  V  T  A  G  L  V  I  P  L  V  P  K  T
 122 CAAGAAGCTATCACCGAGTATGTGAACTCAAAGCAATCTCTCTGGAAGGCTGAGATTCCA 181
      Q  E  A  I  T  E  Y  V  N  S  K  Q  S  L  W  K  A  E  I  P
 182 AAGGACATCACCATTGAACAGGTCAAGAAGCGTCTCATGAGAACCGAATTCGTTGCCCCA 241
      K  D  I  T  I  E  Q  V  K  K  R  L  M  R  T  E  F  V  A  P
 242 CACACTCCAGATGTTGAGGTTGTGAAGCATGATATCAACGAGGATACCATTCCAGCAACA 301
      H  T  P  D  V  E  V  V  K  H  D  I  N  E  D  T  I  P  A  T
 302 TTCGATGCCCGTACCCAATGGCCAAACTGTATGTCAATCAACAACATCCGTGACCAATCT 361
      F  D  A  R  T  Q  W  P  N  C  M  S  I  N  N  I  R  D  Q  S
 362 GACTGTGGATCCTGCTGGGCGTTCGCTGCCGCCGAGGCTGCTTCTGATCGTTTCTGCATC 421
      D  C  G  S  C  W  A  F  A  A  A  E  A  A  S  D  R  F  C  I
 422 GCCTCCAACGGAGCCGTCAACACCCTTCTCTCAGCTGAAGATGTTCTTTCTTGCTGCTCC 481
      A  S  N  G  A  V  N  T  L  L  S  A  E  D  V  L  S  C  C  S
 482 AACTGCGGATACGGATGCGAGGGAGGATACCCAATCAACGCCTGGAAGTACCTTGTCAAG 541
      N  C  G  Y  G  C  E  G  G  Y  P  I  N  A  W  K  Y  L  V  K
 542 AGCGGATTCTGCACCGGAGGATCTTACGAGGCTCAGTTCGGATGCAAGCCATACTCCCTT 601
      S  G  F  C  T  G  G  S  Y  E  A  Q  F  G  C  K  P  Y  S  L
 602 GCCCCATGCGGAGAGACCGTCGGAAACGTTACCTGGCCATCTTGCCCAGATGATGGATAC 661
      A  P  C  G  E  T  V  G  N  V  T  W  P  S  C  P  D  D  G  Y
 662 GATACCCCAGCTTGTGTTAACAAGTGCACCAACAAGAACTACAACGTTGCCTACACCGCT 721
      D  T  P  A  C  V  N  K  C  T  N  K  N  Y  N  V  A  Y  T  A
 722 GACAAGCACTTCGGAAGCACCGCTTACGCCGTCGGAAAGAAGGTCTCCCAGATCCAAGCT 781
      D  K  H  F  G  S  T  A  Y  A  V  G  K  K  V  S  Q  I  Q  A
 782 GAAATCATTGCCCACGGACCAGTCGAGGCCGCATTCACTGTCTACGAGGACTTCTACCAA 841
      E  I  I  A  H  G  P  V  E  A  A  F  T  V  Y  E  D  F  Y  Q
 842 TACAAGACCGGAGTCTACGTTCACACCACTGGACAAGAACTCGGAGGACATGCCATCAGA 901
      Y  K  T  G  V  Y  V  H  T  T  G  Q  E  L  G  G  H  A  I  R
 902 ATTCTTGGATGGGGAACTGACAACGGAACTCCATACTGGCTTGTTGCCAACTCATGGAAC 961
      I  L  G  W  G  T  D  N  G  T  P  Y  W  L  V  A  N  S  W  N
 962 GTCAACTGGGGAGAGAACGGATATTTCCGTATCATCCGTGGAACCAACGAGTGCGGAATT 1021
      V  N  W  G  E  N  G  Y  F  R  I  I  R  G  T  N  E  C  G  I
1022 GAGCACGCCGTTGTCGGAGGAGTCCCAAAAGTCTAAataatctatttgattgaagtattt 1081
      E  H  A  V  V  G  G  V  P  K  V  *
1082 gttttcatacgtgtatatagatatgaataataataaatgattatgcatcgagagtcaaat 1141
1142 gttgattgatatatttgaatcgtggactgataggatcatgaagcacatatttgtgaggag 1201
1202 aaagtttaaacttcgcattttgcttcaatgatcagtttcattaaagtttttcgaagcaaa 1261
1262 atgagaa 1268
```

Figure 4.13

The annotated DNA sequence of *cpr-5*

The coding DNA sequence is indicated in capitals (`CAPITALS`). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (▨). Non-coding sequences (introns and 5′ and 3′ flank sequences) are indicated in lowercase (`lowercase`). The 5′ end of each sequenced 5′ RACE clone is indicated by a vertical arrow (↓). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. The double vertical lines (‖) indicate the start of the poly(A) tail and the boxed nucleotides (▭) indicate the putative cleavage/poly(A) signal. The non-coding nucleotide residues marked in bold (**`bold`**) indicate the region of the 5′ flank of cpr-5 which was only sequenced on one strand.

```
 -324 cgcttatgcttatgcttatgcttttgcttatgcttatgcttatgcttatgcttaggctca -265
 -264 ggcttaggcttaggcattaagcttaggtattaggctcatccctaattcctaatcccctta -205
 -204 gaatttttacagtattaaaaaatgtttacatcaatgatctattccgacaagtgaaagtat -145
 -144 atttggcgcgcgactcatttacgtcacattccttttaattttaattcttaaaaaagata -85
                                                             1
                                                             ↓
  -84 ctgataaaattaactgataagaatgatgcgtaccgactacttaatgagcattagacgcga -25
         1   11 1
         ↓   ↓↓ ↓
  -24 accttcgcagacacttctctcataATGTGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCT 36
                              M  W  K  L  S  A  I  L  L  V  A  A
   37 GCCTCGGCTGTGGTGATTCCTGGACACCGTGAAGCCCCAGCTCTCACTGGACAAGCTTTG 96
      A  S  A  V  V  I  P  G  H  R  E  A  P  A  L  T  G  Q  A  L
   97 ATCGACTATGTCAACTCTGCCCAAAAGCTTTGGACCGCTGGACATCAAGTCATTCCAAAG 156
      I  D  Y  V  N  S  A  Q  K  L  W  T  A  G  H  Q  V  I  P  K
  157 GAGAAGATCACCAAGAAGCTGATGGATGTTAAGTATTTGGTGCCACACAAGGATGAGGAT 216
      E  K  I  T  K  K  L  M  D  V  K  Y  L  V  P  H  K  D  E  D
  217 ATTGTCGCTACTGAGGTCTCCGACGCTATTCCAGACCACTTCGATGCTCGTGATCAATGG 276
      I  V  A  T  E  V  S  D  A  I  P  D  H  F  D  A  R  D  Q  W
  277 CCAAACTGCATGTCCATCAATAACATTAGAGATCAATCTGACTGCGGTTCCTGCTGGGCG 336
      P  N  C  M  S  I  N  N  I  R  D  Q  S  D  C  G  S  C  W  A
  337 TTCGCCGCTGCTGAAGCTATCTCCGACAGAACATGCATTGCCTCCAATGGAGCCGTTAAC 396
      F  A  A  A  E  A  I  S  D  R  T  C  I  A  S  N  G  A  V  N
  397 ACTTTGCTCTCATCTGAAGATCTGCTCTCGTGCTGCACCGGAATGTTCAGCTGCGGAAAT 456
      T  L  L  S  S  E  D  L  L  S  C  C  T  G  M  F  S  C  G  N
  457 GGgtaaggctccgattctgatttgaaaactacgcgtctacacaaaaacgataagaatcga 516
      G
  517 caaattccgtttatgacaaaatgatctctagcgccgtagttttaatctccaaaatcgaaa 576
  577 gttcttgctgataagaacacagtgatcggcatggactcacagttgctttcaaaccattt 636
  637 ttttccagTTGCGAAGGAGGTTACCCAATCCAGGCATGGAAGTGGTGGGTCAAGCACGGT 696
               C  E  G  G  Y  P  I  Q  A  W  K  W  W  V  K  H  G
  697 CTCGTCACCGGAGGATCCTACGAGACCCAGTTCGGATGCAAGCCATACTCGATCGCTCCA 756
      L  V  T  G  G  S  Y  E  T  Q  F  G  C  K  P  Y  S  I  A  P
  757 TGCGGCGAGACTGTGAACGGCGTCAAGTGGCCAGCATGCCCAGAAGACACCGAGCCAACT 816
      C  G  E  T  V  N  G  V  K  W  P  A  C  P  E  D  T  E  P  T
  817 CCAAAGTGTGTCGACTCCTGCACTTCCAAGAACAACTACGCCACCCCATATCTTCAGGAT 876
      P  K  C  V  D  S  C  T  S  K  N  N  Y  A  T  P  Y  L  Q  D
  877 AAGCACTTTGgtgagtttacacgtgtaagcggcgctggttaacactatgaagttctagGA 936
      K  H  F  G
  937 TCTACCGCCTACGCCGTCGGAAAGAAGGTCGAGCAAATCCAGACCGAGATTCTTACGAAC 996
      S  T  A  Y  A  V  G  K  K  V  E  Q  I  Q  T  E  I  L  T  N
  997 GGACCAATCGAGGTTGCCTTCACCGTCTACGAGGACTTCTACCAATACACCACTGGAGTG 1056
      G  P  I  E  V  A  F  T  V  Y  E  D  F  Y  Q  Y  T  T  G  V
 1057 TACGTACACACCGCCGGAGCCTCCCTCGGAGGACACGCCGTCAAGATCCTCGGATGGGGA 1116
      Y  V  H  T  A  G  A  S  L  G  G  H  A  V  K  I  L  G  W  G
 1117 GTCGACAACGGAACCCCATACTGGCTTGTCGCCAACTCGTGGAACGTCGCCTGGGGAGAG 1176
      V  D  N  G  T  P  Y  W  L  V  A  N  S  W  N  V  A  W  G  E
 1177 AAGGGATACTTCCGTATCATTCGTGGACTCAACGAGTGTGGAATCGAGCACTCGGCTGTT 1236
      K  G  Y  F  R  I  I  R  G  L  N  E  C  G  I  E  H  S  A  V
 1237 GCCGGAATTCCAGACTTGGCTCGTCACAACTAAggctgctatcggaactttttttaaaata 1296
      A  G  I  P  D  L  A  R  H  N  *
 1297 attcttcagggagatttccatgtgtatattatgaaatgatgataaatattttataatttg 1356
 1357 ctatgatattatttaaagtataacctcgtttttcagtatctctgtaattggctgaaaaaa 1416
 1417 attggccggtcgaacttttacgagaagggtgggagacggagcggatttgctcaaaaccgg 1476
 1477 cctgacggcagcagacaaaccgcggactagggaacccgtgcaggctgccatgttgcggac 1536
 1537 accttgcggctacctccgcctgccacgcgccgcgccttttgtataatatggctctgaacc 1596
 1597 caaaaagtga 1606
```

Figure 4.14

The annotated DNA sequence of *cpr-6*

The coding DNA sequence is indicated in capitals (CAPITALS). The amino acid residue encoded by each codon is shown below the first nucleotide of each codon. The catalytic amino acid residues, C, H and N, are marked by grey boxes (▒). Non-coding sequences (introns and 5′ and 3′ flank sequences) are indicated in lowercase (`lowercase`). The 5′ end of each sequenced 5′ RACE clone is indicated by a vertical arrow (↓). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. The double vertical lines (‖) indicate the start of the poly(A) tail and the boxed nucleotides (▭) indicate the putative cleavage/poly(A) signal.

```
 -320 attaggttttccatcatataaccctttcaaacgaaattaatgtgctaaatctgttaagt -261
 -260 ttcaatatttccttgtctttaggtcaatcttctttgccacacagttcaaactactaccg -201
 -200 ccgagtcacgtcacaccatcacaggatagtgaccggtcctaggatgtaccctgacactgt -141
 -140 gatggacgcagccgacactcttatcgaaatgcacagggccaaatttgataacgaaaacat -81
                                        131              1
                                        ↓↓↓              ↓
  -80 gttctataaaagcatgctgataaaagcgagcagtcaagcgacgacaacttgcgatcaaca -21
  -20 cgctgaccgtcgacgccaacATGgtaggcttttgaactttgaagtaattttagagaaaa 40
                           M
   41 tttgaattctagAAGACGTTGCTCTTCCTTTCCTGCATAGTGGTAGCAGCTTATTGCGCA 100
                   K  T  L  L  F  L  S  C  I  V  V  A  A  Y  C  A
  101 TGCAATGATAACCTTGAGTCCGTTTTGGgtaggttggatattgatgaagctttctgaaaa 160
       C  N  D  N  L  E  S  V  L  D
  161 tttcaatttattaagACAAATATCGCAATCGTGAAATTGACTCAGAAGCAGCTGAGCTTG 220
                       K  Y  R  N  R  E  I  D  S  E  A  A  E  L  D
  221 ACGGAGATGACTTGATCGACTATGTCAATGAAAACCAAAATCTTTGGACGgtaaacttat 280
        G  D  D  L  I  D  Y  V  N  E  N  Q  N  L  W  T
  281 actcacaaaatattcatctaatggattttcagGCTAAGAAACAAAGACGTTTTTCATCG 340
                                        A  K  K  Q  R  R  F  S  S
  341 GTCTACGGAGAGAACGACAAGGCGAAATGGGGATTGATGGGTGTCAACCATGTCAGACTT 400
       V  Y  G  E  N  D  K  A  K  W  G  L  M  G  V  N  H  V  R  L
  401 TCTGTTAAGgtatgcacatcaaatttgaattcggttatttgaaaacgtcaattgttttga 460
       S  V  K
  461 ttgatagacggcttatcacaaaatagaagagaatcagactgaaacatcaggtgatcaagt 520
  521 tatagatagtgatcttatattcaaacagtgcctatcacttcactcacgtgctcaaccatt 580
  581 ccacccaaacagcacttttcagGGCAAACAACACTTGTCCAAGACCAAGGATCTCGATTT 640
                            G  K  Q  H  L  S  K  T  K  D  L  D  L
  641 GGACATTCCAGAAAGCTTTGATTCTCGTGACAATTGGCCAAAATGCGATTCCATCAAGGT 700
       D  I  P  E  S  F  D  S  R  D  N  W  P  K  C  D  S  I  K  V
  701 CATCAGAGACCAGTCAAGCTGTGGATCCTGCTGGGCTTTCGGAGCCGTTGAGGCAATGTC 760
       I  R  D  Q  S  S  C  G  S  C  W  A  F  G  A  V  E  A  M  S
  761 TGATCGTATTTGCATTGCTTCCCATGGAGAACTTCAAGTTACACTTTCCGCTGATGATCT 820
       D  R  I  C  I  A  S  H  G  E  L  Q  V  T  L  S  A  D  D  L
  821 TCTCAGTTGCTGCAAAAGCTGTGGATTCGGATGTAACGGAGGAGATCCATTGGCTGCCTG 880
       L  S  C  C  K  S  C  G  F  G  C  N  G  G  D  P  L  A  A  W
  881 GCGCTACTGGGTGAAGGATGGAATCGTTACTGGATCAAACTACACCGCTAACAATGGGTG 940
       R  Y  W  V  K  D  G  I  V  T  G  S  N  Y  T  A  N  N  G  C
  941 CAAGgtacaaatagtacaagaataaaagatttcaaactagaacctaacctttttcagCC 1000
       K
                                                                P
 1001 ATACCCATTCCCACCATGTGAGCATCACTCGAAGAAAACCCACTTCGATCCATGTCCACA 1060
       Y  P  F  P  P  C  E  H  H  S  K  K  T  H  F  D  P  C  P  H
 1061 CGATTTGTACCCAACTCCAAAATGTGAAAAGAAGTGCGTTTCTGATTACACTGACAAGAC 1120
       D  L  Y  P  T  P  K  C  E  K  K  C  V  S  D  Y  T  D  K  T
 1121 TTACTCCGAGGACAAATTCTTTGGCGCCAGCGCGTACGGAGTCAAGGATGACGTTGAAGC 1180
       Y  S  E  D  K  F  F  G  A  S  A  Y  G  V  K  D  D  V  E  A
 1181 CATCCAGAAAGAATTGATGACTCACGGACCCCTTGAGATCGCTTTCGAGGTTTACGAGGA 1240
       I  Q  K  E  L  M  T  H  G  P  L  E  I  A  F  E  V  Y  E  D
 1241 TTTCTTGAACTATGACGGTGGAGTCTATGTTgtgagttgtactgttatttgacataaaa 1300
       F  L  N  Y  D  G  G  V  Y  V
 1301 cctgaaaaaaaaattcagCACACCGGAGGAAAGCTCGGAGGAGGACACGCCGTCAAGCTT 1360
                        H  T  G  G  K  L  G  G  G  H  A  V  K  L
 1361 ATCGGATGGGGTATTGACGATGGAATCCCATACTGGACAGTTGCCAACTCTTGGAACACC 1420
       I  G  W  G  I  D  D  G  I  P  Y  W  T  V  A  N  S  W  N  T
 1421 GACTGGGGAGAGGATGGATTCTTCCGTATCCTGAGAGGAGTTGATGAGTGTGGAATTGAA 1480
       D  W  G  E  D  G  F  F  R  I  L  R  G  V  D  E  C  G  I  E
 1481 TCTGGAGTTGTTGGAGGAATTCCAAAGCTCAATAGTCTTACCTCAAGACTTCACAGgtga 1540
       S  G  V  V  G  G  I  P  K  L  N  S  L  T  S  R  L  H  R
 1541 acttttcagctatattgcacgtgacatctaaaaaaaatatgatgtgatttcgtttcatga 1600
 1601 ctcccatgccaatgcccaatttcctaaacggaaacctactttttatctacttaactacta 1660
 1661 aaccaactttttatgtttcagACACCACCGCCGCCACGTCTACGATGACAACTACTGAa 1720
                             H  H  R  R  H  V  Y  D  D  N  Y  *
 1721 ccatcattccatttgaacaaaacctttatttcttttaaatttctatatgtataaaaatga 1780
 1781 atgagttaatcaatatttgcattatagaatgtttctagaagaagttcgtggccgatagaa 1840
 1841 ctttaactgaaatcctaacaacacactaaatcatttgtaattctgcgctcagttccgatt 1900
 1901 gtgtcaaatgttttgcaaagtttcgtgctgttgttctctccggcaatatctttttctt 1960
 1961 tctagaaac 1969
```

Figure 4.15

Schematic representations of the gene architectures of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The diagrams are all drawn to the same scale and the scale is indicated.

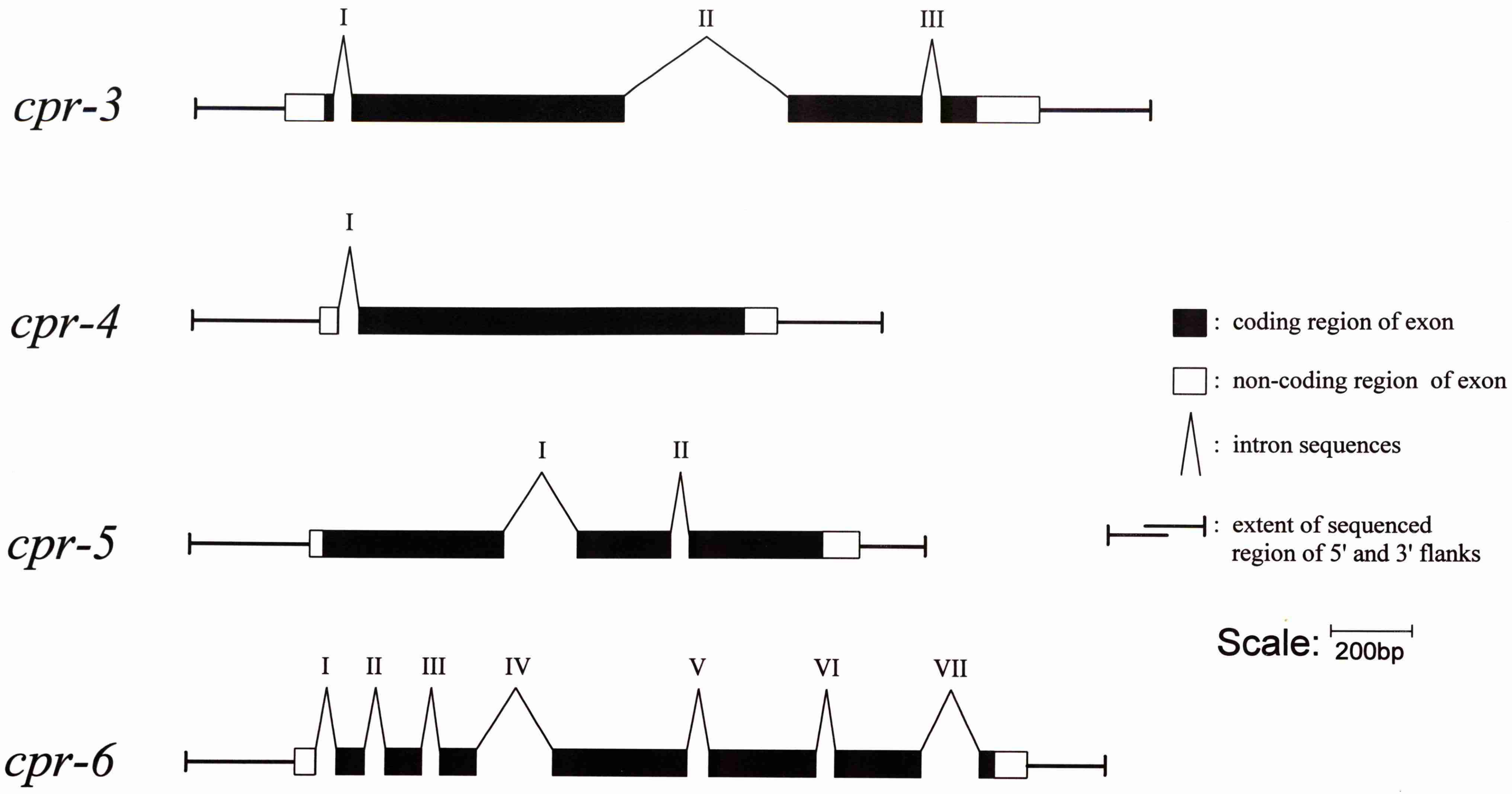

cpr-3
I
II
III
cpr-4
I
cpr-5
I
II
cpr-6
I
II
III
IV
V
VI
VII
: coding region of exon
: non-coding region of exon
: intron sequences
: extent of sequenced region of 5' and 3' flanks
Scale:
200bp

Figure 4.16

Tandemly repeated hexamer sequences within the 5′ flank of *cpr-5*

The figure shows the entire extent of the sequenced region of the 5′ flank of *cpr-5* and a portion of the coding region of the gene. Non-coding DNA sequence is indicated in lowercase (lowercase) and coding sequence in capitals (CAPITALS). The non-coding nucleotide residues marked in bold (**bold**) indicate the region of the 5′ flank of *cpr-5* which was only sequenced on one strand. Single units of the TTATGC hexamer repeat are boxed (ttaggc) and single units of the TTAGGC hexamer repeat are indicated by ovals (ttaggc). Hexamer repeats with single base mismatches to these two repeat sequences are also indicated. These repeats are marked by boxes or ovals, depending on which repeat they most closely resemble. The single base mismatches are underlined (underlined) and the equivalent residue of the appropriate repeat indicated above the mismatch.

```
                            a                                  t
-324 cgcttatgcttatgcttatgcttttgcttatgcttatgcttatgcttatgcttaggctca -265

                        g       c
-264 ggcttaggcttaggcattaagcttaggtattaggctcatccctaattcctaatcccctta -205
-204 gaatttttacagtattaaaaaatgtttacatcaatgatctattccgacaagtgaaagtat -145
-144 atttggcgcgactcatttacgtcacattccttttaatttttaattcttaaaaagata -85
 -84 ctgataaaattaactgataagaatgatgcgtaccgactacttaatgagcattagacgcga -25
 -24 accttcgcagacacttctctcataATGTGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCT +36
                             M  W  K  L  S  A  I  L  L  V  A  A
```

Figure 4.17

Putative regulatory elements in the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The figure shows the DNA sequence of the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* up to the putative ATG translation initiation codon (ATG). The 5′ end of each sequenced 5′ RACE clone is indicated by an arrow (↓). The number above each arrow indicates the number of sequenced 5′ RACE clones whose 5′ ends terminate at that point. Putative regulatory sequence elements are marked in bold (**bold**). Putative TATA promotor elements (TATAWAW) are boxed and underlined (tatawaw). Classic GATA sequence motifs (WGATAR) are indicated by a box with one pointed end, illustrating the orientation of the motif (wgatar). Sequence elements with variations of the classic GATA motif which may bind the GATA family of transcription factors are indicated by an arrow, illustrating their orientation (ngatnn).

cpr-3 5' Flank

```
                                                      -320 tccagacaatttaccctaaa -301
-300 gtgtgaccaaatgcctacattacacacacctcaacgctttgataagcctgataagctccc -241
-240 tcctaattcataatgatgatatcaacgaaggtgataattgatttcttgattcggtgacac -181
-180 taacctccgtgttctcgactgatgactaactttttttttgaatagacaaacataagaggg -121

                                  1
-120 gaaaatgtcgaaattttagtgataagggaaagctacaaaatgtgcggtgtgtgtttgctg -61

     3    3               1
 -60 aaacgcttctttccagttttcagttttccctcaaatttcaaaaattgaatactaaagagaATG +3
```

cpr-4 5' Flank

```
-359  tagtcctcatagttcaaaccttcttgctactttacacctaacctaaaattgagtact -301
-300 tctaatctggttccaaatgataactttcgttgaaccacacaacttccaaactcttatcaa -241
-240 agttgcacgagatcattgtgctcaaatgatggtgctgcgtcacatgactacctcctaatt -181
-180 aggcattgtctatcgaaatttgcgctgccaggtaccgcaaatttttcaattttaatccc -121

-120 ggacgccaatgataagagataacgagcactcccgaactgataacgagtcaactataaaag -61

                           1 4  1
 -60 accatcgcaatgaagtaacttcagcatttgctctatcttgctatttgctcttttacaaaaATG +3
```

cpr-5 5' Flank

```
                                                -324 cgcttatgcttatgcttatgcttt -301
-300 tgcttatgcttatgcttatgcttatgcttaggctcaggcttaggcttaggcattaagctt -241
-240 aggtattaggctcatccctaattcctaatccccttagaatttttacagtattaaaaaatg -181
-180 tttacatcaatgatctattccgacaagtgaaagtatatttggcgcgcgactcatttacgt -121
-120 cacattccttttaatttttaattcttaaaaaagatactgataaaattaactgataaggaat -61

                                     1       11 1
 -60 gatgcgtaccgactacttaatgagcattagacgcgaaccttcgcagacacttctctcataATG +3
```

cpr-6 5' Flank

```
                                                      -320 attaggtttttccatcatat -301
-300 aaccctttcaaacgaaattaatgtgctaaatctgttaagtttcaatattttccttgtctt -241
-240 taggtcaatctttctttgccacacagttcaaactactaccgccgagtcacgtcacaccatc -181

-180 acaggatagtgaccggtcctaggatgtaccctgacactgtgatggacgcagccgacactc -121
-120 ttatcgaaatgcacagggccaaatttgataacgaaaacatgttctataaaagcatgctga -61

                   131            1
 -60 taaaagcgagcagtcaagcgacgacaacttgcgatcaacacgctgaccgtcgacgccaacATG +3
```

Figure 4.18

The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to cysteine protease enzymes in the Genbank and EMBL databases recreated by the NEIGHBOR program

The unrooted phylogenetic tree was generated using the NEIGHBOR program of the PHYLIP PACKAGE version 3.5c (Felsenstein, 1993) and the distances were calculated according to the Dayhoff PAM matrix. The branch lengths represent the percentage of non-synonymous amino acid changes. Node **A** represents the archetype of all cathepsin B and B-like enzymes analysed. Node **B** represents the archetype of all other cysteine protease enzymes analysed.

KEY: Enzyme nomenclature: **cp**, cysteine protease; **cpb**, cathepsin B or B-like cysteine protease; **cps**, cathepsin S cysteine protease; **cpl**, cathepsin L cysteine protease; **cph**, cathepsin H cysteine protease.

Name	Accession Number
hum.cpb	*Homo sapiens* **(M14221)**
moused.cpb	*Mus musculus* **(M14222)**
chicken.cpb	*Gallus gallus* **(U18083)**
bovine.cpb	*Bos taurus* **(L06075)**
sm31.cpb	*Schistosoma mansoni* **(M21309)**
sjapon.cpb	*Schistosoma japonicum* **(X70968)**
spereg.cpb	*Sarcophaga peregrina* **(D16823)**
lmex.cpb	*Leishmania mexicana* **(z48599)**
taesti.cpb	*Triticum aestivum* **(X66013)**
nrusti.cpb	*Nicotiana rustica* **(X81995)**
cecpr1.cpb	*Caenorhabditis elegans, cpr-1* **(M74797)**
cecpr3.cpb	*Caenorhabditis elegans, cpr-3* **(L39890)**
cecpr4.cpb	*Caenorhabditis elegans, cpr-4* **(L39895)**
cecpr5.cpb	*Caenorhabditis elegans, cpr-5* **(L39896)**
cecpr6.cpb	*Caenorhabditis elegans, cpr-6* **(L39894)**
HcAC1.cpb	*Haemonchus contortus*, AC1 **(M31112)**
HcAC2.cpb	*Haemonchus contortus*, AC2 **(M60212, M60213, M34859 and M34860)**
HcAC3.cpb	*Haemonchus contortus*, AC3 **(M80388)**
HcAC4.cpb	*Haemonchus contortus*, AC4 **(M80386)**
HcAC5.cpb	*Haemonchus contortus*, AC5 **(M80385)**
OosterCP1.cpb	*Ostertagia ostertagi*, CP1 **(M88503 and M88504)**
ddisc.cp	*Dictyostelium discoides* **(L36204)**
tthe.cp	*Tetrahymena thermophila* **(L03212)**
lmex.cp	*Leishmania mexicana* **(X62163)**
lpif.cp	*Leishmania pifanoi* **(m97695)**
tbru.cp	*Trypanasoma brucei brucei* **(X16465)**
tcru.cp	*Trypanosoma cruzi* **(X54414)**
tcong1.cp	*Trypanosoma congolense* **(L25130)**
tcong2.cp	*Trypanosoma congolense* **(Z25813)**
tannu.cp	*Theileria annulata* **(M86659)**
tparv1.cp	*Theileria parva* **(M67476)**
tparv2.cp	*Theileria parva* **(M37791)**
pfal.cp	*Plasmodium falciparum* **(P25805)**
pviv.cp	*Plasmodium vivax* **(L26362)**
pvin.cp	*Plasmodium vinckei* **(L08500)**
yeast.cp	*Saccharomyces cerevisiae* **(M97910)**
ehist1.cp	*Entamoeba histolytica* **(M94162)**
ehist2.cp	*Entamoeba histolytica* **(M94163)**
ccarp.cp	*Cyprinus carpio* **(L30111)**
hamer1.cp	*Homarus americanus* **(X63567)**
hamer2.cp	*Homarus americanus* **(X63568)**
hamer3.cp	*Homarus americanus* **(X63569)**
hum.cph	*Homo sapiens* **(P09668)**
hum.cpl	*Homo sapiens* **(P07711)**
hum.cps	*Homo sapiens* **(P25774)**
mouse.cp	*Mus musculus* **(J02583)**
agly.cp	*Alnus glutinosa* **(U13940)**
achd.cp	*Actinidia chinensis* **(P00785)**
barl.cp	barley, Hordeum spp. **(S45163)**
cari.cp	*Cicer arietinum* **(X82011)**
gmax1.cp	*Glycine max* **(Z32795)**
gmax2.cp	*Glycine max* **(D28876)**
osat1.cp	*Oryza sativa* **(P25778)**
osat2.cp	*Oryza sativa* **(X80876)**
vsat1.cp	*Vicia sativa* **(Z30338)**
vsat2.cp	*Vicia sativa* **(Z34895)**
vsat3.cp	*Vicia sativa* **(X75749)**
vsat4.cp	*Vicia sativa* **(Z34899)**
zmays1.cp	*Zea mays* **(D45402)**
zmays2.cp	*Zea mays* **(D45403)**
ntaba1.cp	*Nicotiana tabacum* **(Z13964)**
ntaba2.cp	*Nicotiana tabacum* **(Z13965)**

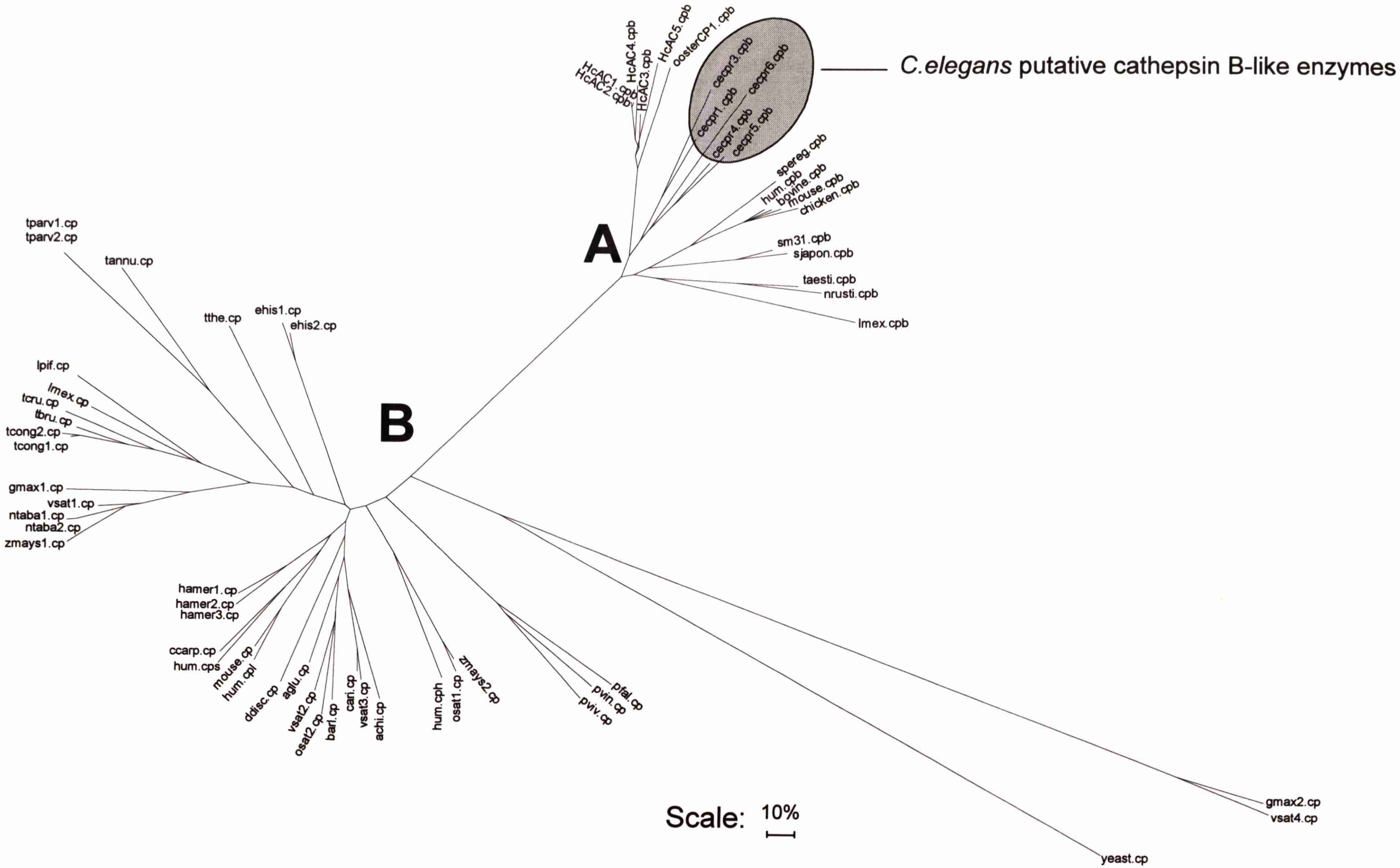

C.elegans putative cathepsin B-like enzymes
A
B
HcAC1.cpb
HcAC2.cpb
HcAC4.cpb
HcAC3.cpb
HcAC5.cpb
oosterCP1.cpb
cecpr3.cpb
cecpr1.cpb
cecpr6.cpb
cecpr4.cpb
cecpr5.cpb
spereg.cpb
hum.cpb
bovine.cpb
mouse.cpb
chicken.cpb
sm31.cpb
sjapon.cpb
taesti.cpb
nrusti.cpb
lmex.cpb
tparv1.cp
tparv2.cp
tannu.cp
tthe.cp
ehis1.cp
ehis2.cp
lpif.cp
lmex.cp
tcru.cp
tbru.cp
tcong2.cp
tcong1.cp
gmax1.cp
vsat1.cp
ntaba1.cp
ntaba2.cp
zmays1.cp
hamer1.cp
hamer2.cp
hamer3.cp
ccarp.cp
hum.cps
mouse.cp
hum.cpl
ddisc.cp
aglu.cp
vsat2.cp
osat2.cp
barl.cp
cari.cp
vsat3.cp
achi.cp
hum.cph
osat1.cp
zmays2.cp
pfal.cp
pvin.cp
pviv.cp
gmax2.cp
vsat4.cp
yeast.cp
Scale: 10%

Figure 4.19

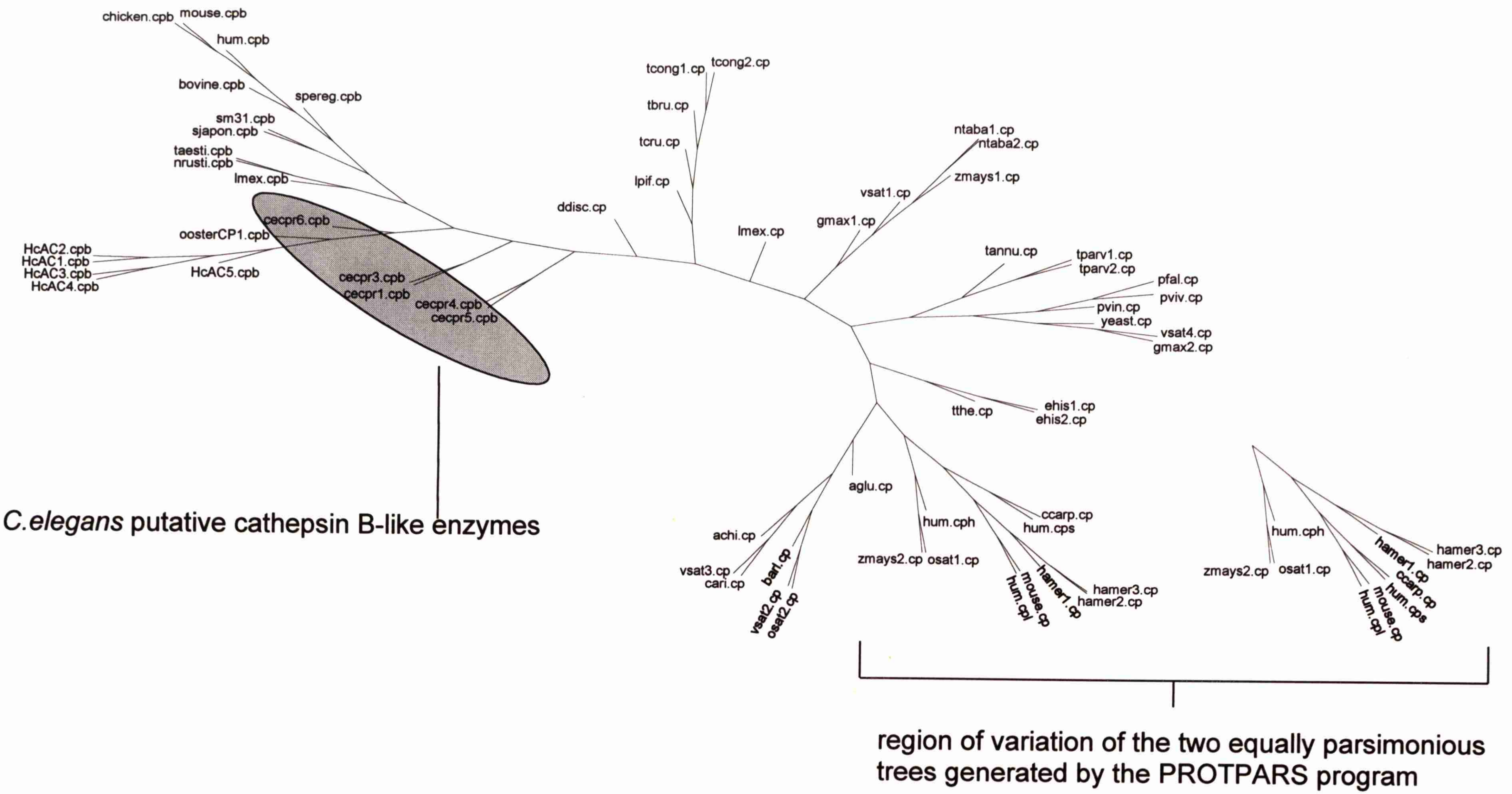
The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to cysteine protease enzymes in the Genbank and EMBL databases, recreated by the PROTPARS program

Two equally parsimonious unrooted phylogenetic trees was generated using the PROTPARS program of the PHYLIP PACKAGE version 3.5c (Felsenstein, 1993). These two trees were identical with the exception of one cluster. Therefore, only one of the trees, and the region of the second tree that differed, is shown.

KEY: Enzyme nomenclature: **cp**, cysteine protease; **cpb**, cathepsin B or B-like cysteine protease; **cps**, cathepsin S cysteine protease; **cpl**, cathepsin L cysteine protease; **cph**, cathepsin H cysteine protease.

Name	Accession Number
hum.cpb	*Homo sapiens* **(M14221)**
moused.cpb	*Mus musculus* **(M14222)**
chicken.cpb	*Gallus gallus* **(U18083)**
bovine.cpb	*Bos taurus* **(L06075)**
sm31.cpb	*Schistosoma mansoni* **(M21309)**
sjapon.cpb	*Schistosoma japonicum* **(X70968)**
spereg.cpb	*Sarcophaga peregrina* **(D16823)**
lmex.cpb	*Leishmania mexicana* **(z48599)**
taesti.cpb	*Triticum aestivum* **(X66013)**
nrusti.cpb	*Nicotiana rustica* **(X81995)**
cecpr1.cpb	*Caenorhabditis elegans, cpr-1* **(M74797)**
cecpr3.cpb	*Caenorhabditis elegans, cpr-3* **(L39890)**
cecpr4.cpb	*Caenorhabditis elegans, cpr-4* **(L39895)**
cecpr5.cpb	*Caenorhabditis elegans, cpr-5* **(L39896)**
cecpr6.cpb	*Caenorhabditis elegans, cpr-6* **(L39894)**
HcAC1.cpb	*Haemonchus contortus*, AC1 **(M31112)**
HcAC2.cpb	*Haemonchus contortus*, AC2 **(M60212, M60213, M34859 and M34860)**
HcAC3.cpb	*Haemonchus contortus*, AC3 **(M80388)**
HcAC4.cpb	*Haemonchus contortus*, AC4 **(M80386)**
HcAC5.cpb	*Haemonchus contortus*, AC5 **(M80385)**
OosterCP1.cpb	*Ostertagia ostertagi*, CP1 **(M88503 and M88504)**
ddisc.cp	*Dictyostelium discoides* **(L36204)**
tthe.cp	*Tetrahymena thermophila* **(L03212)**
lmex.cp	*Leishmania mexicana* **(X62163)**
lpif.cp	*Leishmania pifanoi* **(m97695)**
tbru.cp	*Trypanasoma brucei brucei* **(X16465)**

Name	Accession Number
tcru.cp	*Trypanosoma cruzi* **(X54414)**
tcong1.cp	*Trypanosoma congolense* **(L25130)**
tcong2.cp	*Trypanosoma congolense* **(Z25813)**
tannu.cp	*Theileria annulata* **(M86659)**
tparv1.cp	*Theileria parva* **(M67476)**
tparv2.cp	*Theileria parva* **(M37791)**
pfal.cp	*Plasmodium falciparum* **(P25805)**
pviv.cp	*Plasmodium vivax* **(L26362)**
pvin.cp	*Plasmodium vinckei* **(L08500)**
yeast.cp	*Saccharomyces cerevisiae* **(M97910)**
ehist1.cp	*Entamoeba histolytica* **(M94162)**
ehist2.cp	*Entamoeba histolytica* **(M94163)**
ccarp.cp	*Cyprinus carpio* **(L30111)**
hamer1.cp	*Homarus americanus* **(X63567)**
hamer2.cp	*Homarus americanus* **(X63568)**
hamer3.cp	*Homarus americanus* **(X63569)**
hum.cph	*Homo sapiens* **(P09668)**
hum.cpl	*Homo sapiens* **(P07711)**
hum.cps	*Homo sapiens* **(P25774)**
mouse.cp	*Mus musculus* **(J02583)**
agly.cp	*Alnus glutinosa* **(U13940)**
achd.cp	*Actinidia chinensis* **(P00785)**
barl.cp	barley, Hordeum spp. **(S45163)**
cari.cp	*Cicer arietinum* **(X82011)**
gmax1.cp	*Glycine max* **(Z32795)**
gmax2.cp	*Glycine max* **(D28876)**
osat1.cp	*Oryza sativa* **(P25778)**
osat2.cp	*Oryza sativa* **(X80876)**
vsat1.cp	*Vicia sativa* **(Z30338)**
vsat2.cp	*Vicia sativa* **(Z34895)**
vsat3.cp	*Vicia sativa* **(X75749)**
vsat4.cp	*Vicia sativa* **(Z34899)**
zmays1.cp	*Zea mays* **(D45402)**
zmays2.cp	*Zea mays* **(D45403)**
ntaba1.cp	*Nicotiana tabacum* **(Z13964)**
ntaba2.cp	*Nicotiana tabacum* **(Z13965)**



C.elegans putative cathepsin B-like enzymes

region of variation of the two equally parsimonious trees generated by the PROTPARS program
]

Figure 4.20

The phylogenetic relationships of the putative cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to cathepsin B and B-like enzymes in the Genbank and EMBL databases

A and B). Two different representations of the same unrooted majority-rule consensus phylogenetic tree generated using the PROTPARS program of the PHYLIP package version 3.5c (Felsenstein, 1993). Bootstrap resampling was performed with the PROTPARS program and the majority-rule consensus tree was obtained from the 20 trees generated by the bootstrap analysis using the CONSENSUS program of the PHYLIP package, version 3.5c. In Figure 4.20B, the number by each grouping represents the percentage of the 20 trees generated by Bootstrap resampling that support that particular grouping.

C). Unrooted phylogenetic tree generated using the NEIGHBOR program of the PHYLIP PACKAGE version 3.5c. The distances were calculated according to the Dayhoff PAM matrix. The branch lengths represent the percentage of non-synonymous amino acid changes.

KEY: Enzyme nomenclature: **cpb**, cathepsin B or B-like cysteine protease

The grey elipses highlight the three groups of putative *C.elegans* cathepsin B-like enzymes generay=ted by the NEIGHBOR and PROTPARS programs.

Name	Accession Number	Name	Accession Number
hum.cpb	***Homo sapiens* (M14221)**	cecpr5.cpb	***Caenorhabditis elegans, cpr-5* (L39896)**
moused.cpb	***Mus musculus* (M14222)**	cecpr6.cpb	***Caenorhabditis elegans, cpr-6* (L39894)**
chicken.cpb	***Gallus gallus* (U18083)**	HcAC1.cpb	***Haemonchus contortus*, AC1 (M31112)**
bovine.cpb	***Bos taurus* (L06075)**	HcAC2.cpb	***Haemonchus contortus*, AC2 (M60212, M60213, M34859 and M34860)**
sm31.cpb	***Schistosoma mansoni* (M21309)**	HcAC3.cpb	***Haemonchus contortus*, AC3 (M80388)**
sjapon.cpb	***Schistosoma japonicum* (X70968)**	HcAC4.cpb	***Haemonchus contortus*, AC4 (M80386)**
spereg.cpb	***Sarcophaga peregrina* (D16823)**	HcAC5.cpb	***Haemonchus contortus*, AC5 (M80385)**
lmex.cpb	***Leishmania mexicana* (z48599)**	OosterCP1.cpb	***Ostertagia ostertagi*, CP1 (M88503 and M88504)**
taesti.cpb	***Triticum aestivum* (X66013)**		
nrusti.cpb	***Nicotiana rustica* (X81995)**		
cecpr1.cpb	***Caenorhabditis elegans, cpr-1* (M74797)**		
cecpr3.cpb	***Caenorhabditis elegans, cpr-3* (L39890)**		
cecpr4.cpb	***Caenorhabditis elegans, cpr-4* (L39895)**		

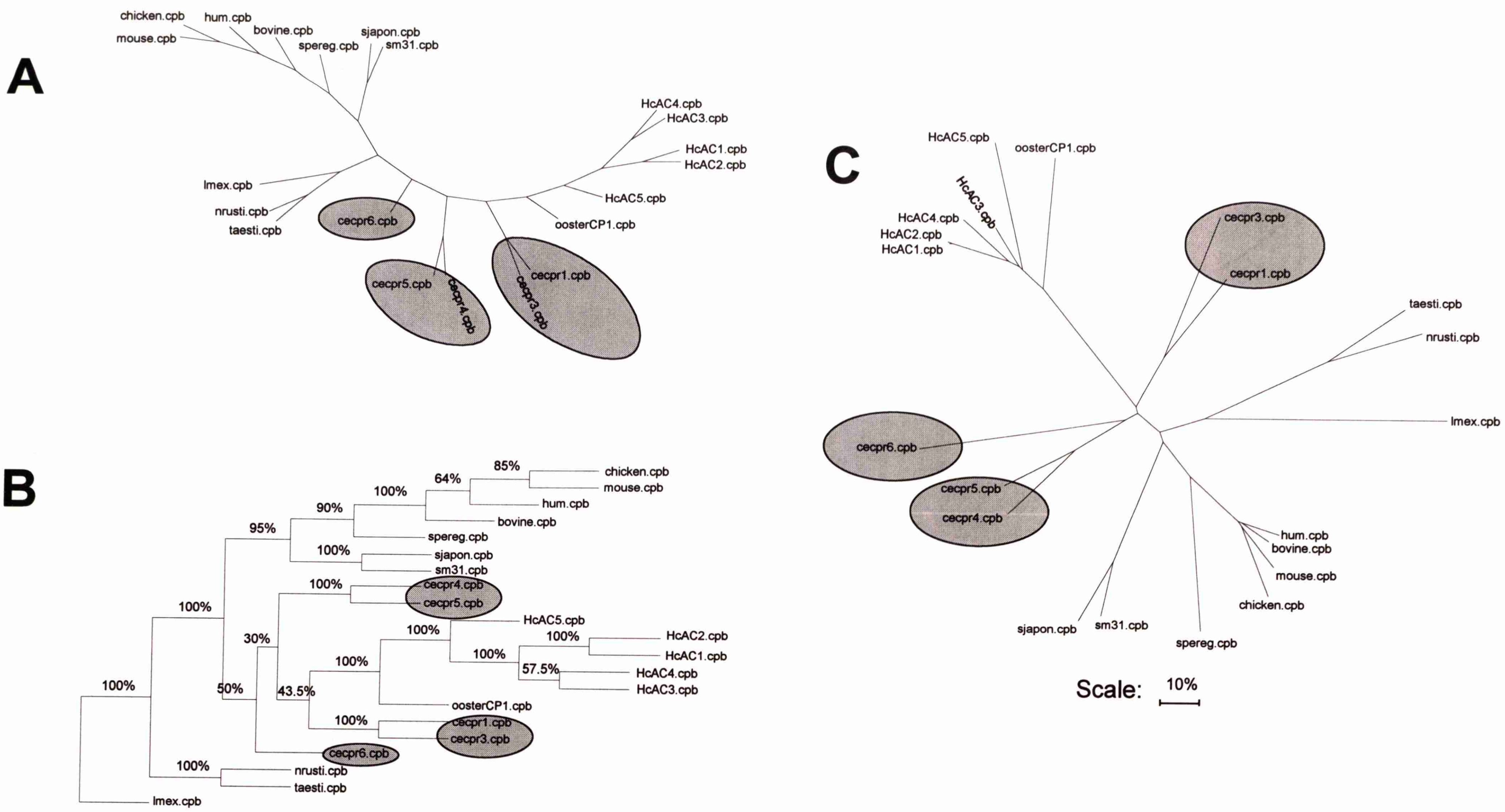

A
chicken.cpb
mouse.cpb
hum.cpb
bovine.cpb
spereg.cpb
sjapon.cpb
sm31.cpb
HcAC4.cpb
HcAC3.cpb
HcAC1.cpb
HcAC2.cpb
HcAC5.cpb
oosterCP1.cpb
lmex.cpb
nrusti.cpb
taesti.cpb
cecpr6.cpb
cecpr5.cpb
cecpr4.cpb
cecpr1.cpb
cecpr3.cpb
B
85%
64%
100%
90%
95%
100%
100%
100%
30%
100%
100%
100%
100%
57.5%
100%
50%
43.5%
100%
100%
chicken.cpb
mouse.cpb
hum.cpb
bovine.cpb
spereg.cpb
sjapon.cpb
sm31.cpb
cecpr4.cpb
cecpr5.cpb
HcAC5.cpb
HcAC2.cpb
HcAC1.cpb
HcAC4.cpb
HcAC3.cpb
oosterCP1.cpb
cecpr1.cpb
cecpr3.cpb
cecpr6.cpb
nrusti.cpb
taesti.cpb
lmex.cpb
C
HcAC5.cpb
oosterCP1.cpb
HcAC3.cpb
HcAC4.cpb
HcAC2.cpb
HcAC1.cpb
cecpr3.cpb
cecpr1.cpb
taesti.cpb
nrusti.cpb
lmex.cpb
cecpr6.cpb
cecpr5.cpb
cecpr4.cpb
hum.cpb
bovine.cpb
mouse.cpb
chicken.cpb
sjapon.cpb
sm31.cpb
spereg.cpb
Scale:
10%

Figure 4.21

An alignment of the predicted amino acid sequences of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5*, *cpr-6*, AC-2, CP-1 and mouse cathepsin B

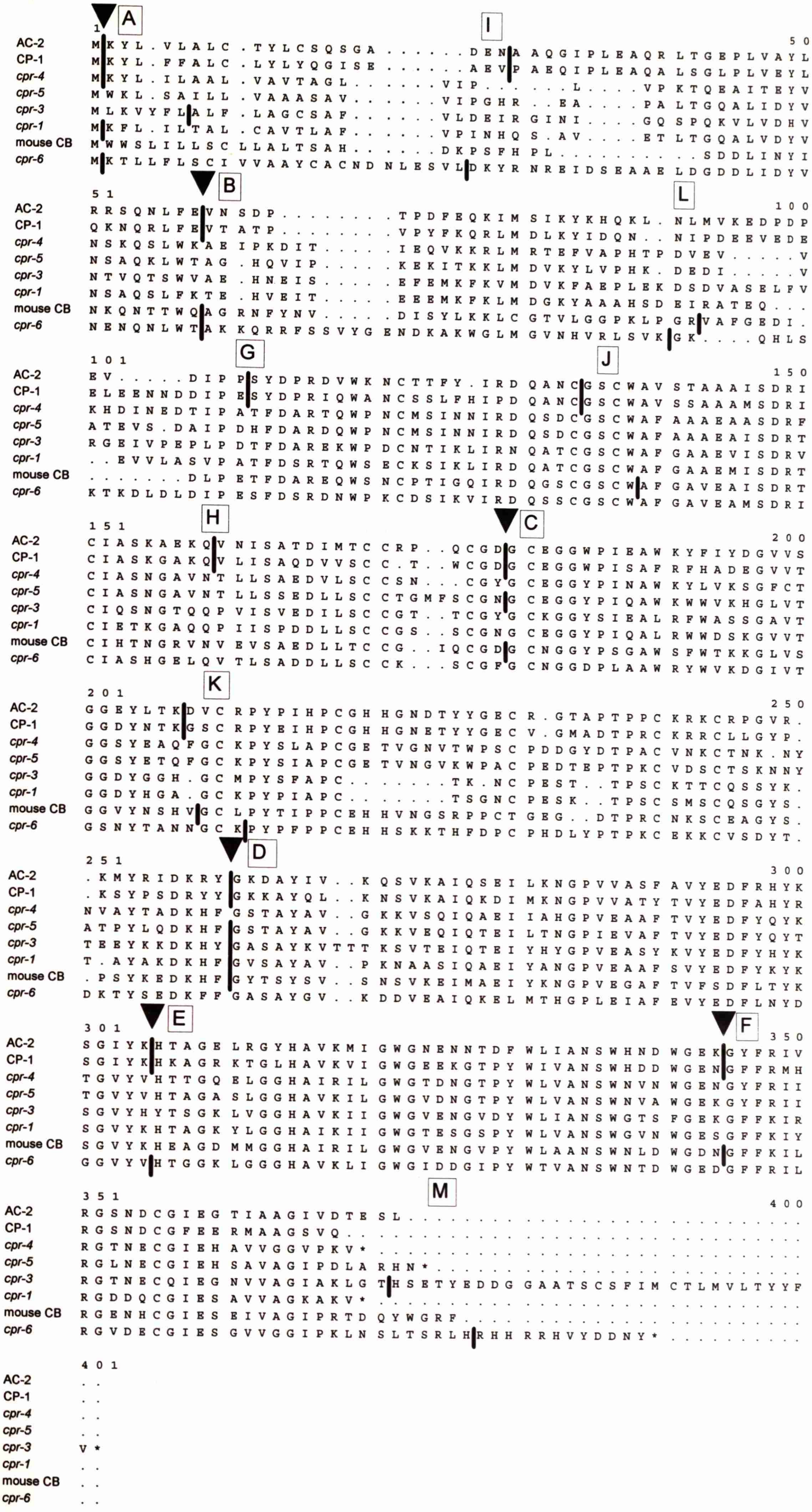
The alignment was generated using the PILEUP program of the GCG package. Vertical bars (**|**) mark intron positions with respect to the amino acid sequence. *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encode the cathepsin B-like genes of *C.elegans*. AC-2 encodes a cathepsin B-like gene from *H.contortus*. CP-1 encodes a cathepsin B-like gene from *O.ostertagi*.

A-F denote intron positions conserved between genes that are not only conserved with respect to the amino acid sequence but also to the exact nucleotides at which intron / exon boundaries occur within codons. These positions are also marked by an arrow (▼). For AC-2 and CP-1, all 11 intron positions show this level of conservation.

H denotes a site of possible intron loss

I-K denote sites where introns are found in close proximity to the intron sites defined by AC-2 and CP-1 but are not at identical positions

L and M denote sites where introns are found in close proximity to one another but outwith the intron sites defined by AC-2 and CP-1



```
           1                                                 50
AC-2       MKYL.VLALC .TYLCSQSGA ......DENA AQGIPLEAQR LTGEPLVAYL
CP-1       MKYL.FFALC .LYLYQGISE ......AEVP AEQIPLEAQA LSGLPLVEYL
cpr-4      MKYL.ILAAL .VAVTAGL.. ....VIP... ...L....VP KTQEAITEYV
cpr-5      MWKL.SAILL .VAAASAV.. ....VIPGHR ..EA....PA LTGQALIDYV
cpr-3      MLKVYFLALF .LAGCSAF.. ....VLDEIR GINI....GQ SPQKVLVDHV
cpr-1      MKFL.ILTAL .CAVTLAF.. ....VPINHQ S.AV....ET LTGQALVDYV
mouse CB   MWWSLILLSC LLALTSAH.. ....DKPSFH PL........ ..SDDLINYI
cpr-6      MKTLLFLSCI VVAAYCACND NLESVLDKYR NREIDSEAAE LDGDDLIDYV

           51                                               100
AC-2       RRSQNLFEVN SDP....... .TPDFEQKIM SIKYKHQKL. NLMVKEDPDP
CP-1       QKNQRLFEVT ATP....... .VPYFKQRLM DLKYIDQN.. NIPDEEVEDE
cpr-4      NSKQSLWKAE IPKDIT.... .IEQVKKRLM RTEFVAPHTP DVEV.....V
cpr-5      NSAQKLWTAG .HQVIP.... .KEKITKKLM DVKYLVPHK. DEDI.....V
cpr-3      NTVQTSWVAE .HNEIS.... .EFEMKFKVM DVKFAEPLEK DSDVASELFV
cpr-1      NSAQSLFKTE .HVEIT.... .EEEMKFKLM DGKYAAAHSD EIRATEQ...
mouse CB   NKQNTTWQAG RNFYNV.... .DISYLKKLC GTVLGGPKLP GRVAFGEDI.
cpr-6      NENQNLWTAK KQRRFSSVYG ENDKAKWGLM GVNHVRLSVK GK....QHLS

           101                                              150
AC-2       EV.....DIP PSYDPRDVWK NCTTFY.IRD QANCGSCWAV STAAAISDRI
CP-1       ELEENNDDIP ESYDPRIQWA NCSSLFHIPD QANCGSCWAV SSAAAMSDRI
cpr-4      KHDINEDTIP ATFDARTQWP NCMSINNIRD QSDCGSCWAF AAAEAASDRF
cpr-5      ATEVS.DAIP DHFDARDQWP NCMSINNIRD QSDCGSCWAF AAAEAISDRT
cpr-3      RGEIVPEPLP DTFDAREKWP DCNTIKLIRN QATCGSCWAF GAAEVISDRV
cpr-1      ..EVVLASVP ATFDSRTQWS ECKSIKLIRD QATCGSCWAF GAAEMISDRT
mouse CB   .......DLP ETFDAREQWS NCPTIGQIRD QGSCGSCWAF GAVEAISDRT
cpr-6      KTKDLDLDIP ESFDSRDNWP KCDSIKVIRD QSSCGSCWAF GAVEAMSDRI

           151                                              200
AC-2       CIASKAEKQV NISATDIMTC CRP..QCGDG CEGGWPIEAW KYFIYDGVVS
CP-1       CIASKGAKQV LISAQDVVSC C.T..WCGDG CEGGWPISAF RFHADEGVVT
cpr-4      CIASNGAVNT LLSAEDVLSC CSN...CGYG CEGGYPINAW KYLVKSGFCT
cpr-5      CIASNGAVNT LLSSEDLLSC CTGMFSCGNG CEGGYPIQAW KWWVKHGLVT
cpr-3      CIQSNGTQQP VISVEDILSC CGT..TCGYG CKGGYSIEAL RFWASSGAVT
cpr-1      CIETKGAQQP IISPDDLLSC CGS..SCGNG CEGGYPIQAL RWWDSKGVVT
mouse CB   CIHTNGRVNV EVSAEDLLTC CG..IQCGDG CNGGYPSGAW SFWTKKGLVS
cpr-6      CIASHGELQV TLSADDLLSC CK...SCGFG CNGGDPLAAW RYWVKDGIVT

           201                                              250
AC-2       GGEYLTKDVC RPYPIHPCGH HGNDTYYGEC R.GTAPTPPC KRKCRPGVR.
CP-1       GGDYNTKGSC RPYEIHPCGH HGNETYYGEC V.GMADTPRC KRRCLLGYP.
cpr-4      GGSYEAQFGC KPYSLAPCGE TVGNVTWPSC PDDGYDTPAC VNKCTNK.NY
cpr-5      GGSYETQFGC KPYSIAPCGE TVNGVKWPAC PEDTEPTPKC VDSCTSKNNY
cpr-3      GGDYGGH.GC MPYSFAPC.. .....TK.NC PEST..TPSC KTTCQSSYK.
cpr-1      GGDYHGA.GC KPYPIAPC.. .....TSGNC PESK..TPSC SMSCQSGYS.
mouse CB   GGVYNSHVGC LPYTIPPCEH HVNGSRPPCT GEG..DTPRC NKSCEAGYS.
cpr-6      GSNYTANNGC KPYPFPPCEH HSKKTHFDPC PHDLYPTPKC EKKCVSDYT.

           251                                              300
AC-2       .KMYRIDKRY GKDAYIV..K QSVKAIQSEI LKNGPVVASF AVYEDFRHYK
CP-1       .KSYPSDRYY GKKAYQL..K NSVKAIQKDI MKNGPVVATY TVYEDFAHYR
cpr-4      NVAYTADKHF GSTAYAV..G KKVSQIQAEI IAHGPVEAAF TVYEDFYQYK
cpr-5      ATPYLQDKHF GSTAYAV..G KKVEQIQTEI LTNGPIEVAF TVYEDFYQYT
cpr-3      TEEYKKDKHY GASAYKVTTT KSVTEIQTEI YHYGPVEASY KVYEDFYHYK
cpr-1      T.AYAKDKHF GVSAYAV..P KNAASIQAEI YANGPVEAAF SVYEDFYKYK
mouse CB   .PSYKEDKHF GYTSYSV..S NSVKEIMAEI YKNGPVEGAF TVFSDFLTYK
cpr-6      DKTYSEDKFF GASAYGV..K DDVEAIQKEL MTHGPLEIAF EVYEDFLNYD

           301                                              350
AC-2       SGIYKHTAGE LRGYHAVKMI GWGNENNTDF WLIANSWHND WGEKGYFRIV
CP-1       SGIYKHKAGR KTGLHAVKVI GWGEEKGTPY WIVANSWHDD WGENGFFRMH
cpr-4      TGVYVHTTGQ ELGGHAIRIL GWGTDNGTPY WLVANSWNVN WGENGYFRII
cpr-5      TGVYVHTAGA SLGGHAVKIL GWGVDNGTPY WLVANSWNVA WGEKGYFRII
cpr-3      SGVYHYTSGK LVGGHAVKII GWGVENGVDY WLIANSWGTS FGEKGFFKIR
cpr-1      SGVYKHTAGK YLGGHAIKII GWGTESGSPY WLVANSWGVN WGESGFFKIY
mouse CB   SGVYKHEAGD MMGGHAIRIL GWGVENGVPY WLAANSWNLD WGDNGFFKIL
cpr-6      GGVYVHTGGK LGGGHAVKLI GWGIDDGIPY WTVANSWNTD WGEDGFFRIL

           351                                              400
AC-2       RGSNDCGIEG TIAAGIVDTE SL........ .......... ..........
CP-1       RGSNDCGFEE RMAAGSVQ.. .......... .......... ..........
cpr-4      RGTNECGIEH AVVGGVPKV* .......... .......... ..........
cpr-5      RGLNECGIEH SAVAGIPDLA RHN*...... .......... ..........
cpr-3      RGTNECQIEG NVVAGIAKLG THSETYEDDG GAATSCSFIM CTLMVLTYYF
cpr-1      RGDDQCGIES AVVAGKAKV* .......... .......... ..........
mouse CB   RGENHCGIES EIVAGIPRTD QYWGRF.... .......... ..........
cpr-6      RGVDECGIES GVVGGIPKLN SLTSRLHRHH RRHVYDDNY* ..........

           401
AC-2       ..
CP-1       ..
cpr-4      ..
cpr-5      ..
cpr-3      V*
cpr-1      ..
mouse CB   ..
cpr-6      ..
```

Chapter 5

Chapter 5

The Temporal and Spatial Expression Patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

5.1. Introduction

In order to further characterise *cpr-3*, *cpr-4*, *cpr-5* and *cpr*-6, I decided to analyse their temporal and spatial patterns of expression. The temporal and spatial expression pattern of a gene can provide a substantial amount of information regarding its biological function. For example, expression restricted to certain cell types (or tissues) may indicate a specific function for that gene in those cells or tissues. In such circumstances, the biological roles associated with a particular cell or tissue type may help assign a potential biological role for the gene. By analogy, it may also be possible to infer the biological role of a gene from its temporal expression pattern. For example, a gene which is expressed within a defined period of time during embryogenesis may be required for embryogenesis. When such expression patterns are matched with the DNA sequence of the gene analysed, it may be possible to accurately predict the biological function of that gene. For example, DNA sequencing might reveal the isolation of a collagen gene, while expression analysis might reveal that the gene is expressed cyclically in the hypodermis prior to each larval moult. Together, such information would provide strong evidence of a structural role for this gene in the *C.elegans* cuticle, which could not be obtained from the DNA sequence or expression data alone.

When considering gene families, the level of complexity is much increased since the biological functions performed by the gene family may require the interaction of none, some or all members of the gene family to varying degrees. Thus, at one extreme, all members of a gene family may be required to act in concert in order to perform their biological roles while, at the other extreme, all members of a gene family may act independently of one another to perform distinct biological roles. Together, determining the DNA sequence of the members comprising a gene family as well as determining their spatial and temporal expression patterns may help to determine the degree of interaction between the various genes comprising the gene family. The results obtained so far for the cathepsin B-like gene family of *C.elegans* provide a good example for illustrating this. From the results discussed in Chapters 3 and 4, there is a substantial amount of evidence to suggest that the cathepsin B-like genes of *C.elegans* are highly diverged and encode enzymes with different substrate specificities and activities. Accordingly, the discovery that these genes are expressed in the same tissue but at different stages during *C.elegans* development would suggest a requirement for different enzymes, or combinations of enzymes, to process or degrade different substrates, or combinations of substrates, at different stages during *C.elegans* development. Alternatively, the discovery that all these genes are expressed in the same tissue and at the same time would suggest that the enzymes may be required at the same time to effectively broaden the spectrum of substrates which can be digested or processed at any one time. In both these examples, different members of the gene family may have to interact substantially in order to perform their functions. For example, a particular substrate might require the activity of several different isoenzymes for its complete digestion or processing. In other scenarios, the genes may be expressed in different tissues, either at the same time or at different times. Such situations may reflect a requirement for slightly different enzymes in different cellular environments which raises the possibility that these genes may have different biological roles by virtue of their involvement in distinct cellular processes.

None of the hypothetical situations, including those discussed above, are mutually exclusive. Accordingly, large gene families may exhibit complex temporal and spatial expression patterns, with different members of a gene family interacting with each other to different extents. Comparison of different expression patterns may allow the members comprising a gene family to be ordered into subgroups on the basis of the frequency with which members of the same group are expressed at the same time in the same tissues, or on the basis of which members are never expressed together. Such analysis may be extremely useful for understanding how members of a gene family interact. For example, there is now evidence to suggest that, though cuticular collagen genes are expressed in the hypodermis, different groups of these genes are expressed at slightly different times prior to moulting (I.L.Johnstone, pers. comm.). It is thought that those collagen genes expressed simultaneously are capable of producing functional heterotrimeric proteins required for cuticle formation, but are not capable of doing so with collagens expressed at different stages. This may therefore represent a mechanism by which incompatible members of a gene family are effectively isolated from one another.

To conclude, I decided to analyse the temporal and spatial expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr*-6. Such studies, in conjunction with the DNA sequence data for the four genes, would provide information essential for understanding both the biological roles and biological interactions of the protein products encoded by these genes.

5.2. Analysing the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr*-6

5.2.1. Experimental approaches used

I decided to use Northern blot analysis and semi-quantitative reverse transcription PCR (s-q rtPCR), to study the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr*-6. Northern blot analysis is one of the simplest and most widely used approaches for assaying the temporal expression patterns of genes during development. This approach is known to be reliable and standard protocols for this method are readily available for *C.elegans*. Developmental Northern blot analysis requires the availability of cloned genes that are constitutively expressed during the life cycle. These clones can be used to control for equal loading in different lanes of RNA isolated from different developmental stages, allowing comparisons of transcript abundance to be made between samples. Such cloned genes are available for use as controls for both total RNA and poly(A)$^+$ selected RNA extracted from *C.elegans*. The pCe7 clone, which contains a single repeat of the 18s and 26s ribosomal DNA sequences (Files and Hirsh, 1981), and the pCeIF clone, which contains a cDNA encoding the *C.elegans* homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992), are recommended for use as loading controls for total RNA and poly(A)$^+$ selected RNA, respectively.

Northern blot analysis is relatively insensitive when compared to other approaches. The lack of sensitivity necessitates generating large synchronous populations of *C.elegans* in order to obtain sufficient material for analysis from each developmental stage. Though the generation of synchronous populations of *C.elegans* can be achieved relatively easily by the isolation of eggs (Sulston and Hodgkin, 1988) and utilisation of the L_1 arrest caused by starving, the degree of synchrony decreases dramatically with each subsequent generation. Thus, large numbers of eggs must be obtained initially in order to generate a sufficiently large synchronous population such that all the desired RNA samples can be extracted within one generation. Since the culture of extremely large populations of *C.elegans* is unfeasible in the laboratory, there is an upper limit on the total number of eggs that can be obtained in order to generate a synchronous population. Furthermore, for reasons which remain unclear, we have found that better synchrony is obtained with small populations rather than with large, dense cultures. These factors limit the maximum size of the synchronous population which can be generated. In addition to the requirement for large amounts of material, the insensitivity of Northern blot analysis also sets a limit on the frequency with which a synchronous population can be sampled, since sampling too frequently will result in insufficient material for analysis. A high frequency of sampling is essential if temporal patterns of expression of a target gene are to be accurately assayed. If sampling becomes too infrequent, or the sample window too large, transcripts from genes with very restricted temporal patterns of expression may be missed, giving rise to inaccurate results. For these reasons I decided to use Northern blot analysis to assay the transcript abundance of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* from only two samples, embryonic RNA and mixed stage RNA. This analysis would give some insight into the temporal pattern of expression of each of these genes. More importantly, it would reveal both the sizes of the transcripts generated and whether or not each of these genes generates multiple mRNA size species.

Semi-quantitative reverse transcriptase polymerase chain reaction (s-q rtPCR) was also used to assay the temporal expression patterns of the four genes. This approach incorporates a PCR amplification step which makes it much more sensitive than Northern blot analysis, however the technique is also inherently more complex.

In brief, the method I used involves converting RNA from each sample to cDNA, using reverse transcriptase. Two sets of PCR primers; one set, specific to the target transcript and the other, to a constitutively expressed transcript acting as an internal control, are then used to amplify the cDNAs simultaneously in a single reaction. The number of PCR cycles is chosen empirically to ensure that reactants are still in excess after amplification, and do not therefore become rate limiting. After amplification, the

PCR products are electrophoresed through an agarose gel and Southern blotted to a nylon filter. The filter is probed using sequences specific to the target and control genes. The resulting autoradiograph is subsequently used as a template to excise the regions of the filter corresponding to the amplification products of the cDNA molecules derived from the target and control transcripts. The amount of radioactivity from each band is then assayed by counting in scintillant. The abundance of the target gene can therefore be expressed as a ratio of the signal obtained from the target gene to that of the control gene. This method therefore measures the relative abundance of a target transcript with respect to a previously chosen constitutively expressed reference transcript as an internal control. The method gives no indication of the absolute abundance of any given transcript. However, it can be used to plot fluctuations of that transcript's abundance in different samples of RNA, for example a time course, by comparing the test gene to internal control gene relative abundance ratio between samples.

The PCR amplification step responsible for the much increased sensitivity of s-q rtPCR is also a potential source of error, since many factors can affect the efficiency of amplification when using PCR, especially small variations in the relative concentrations of the reactants within each tube. This problem is largely overcome by introducing an internal control transcript for the PCR reaction, since most factors which affect the efficiency of amplification of cDNA derived from the target transcript will also affect the efficiency of amplification of cDNA derived from the control transcript. Accordingly, s-q rtPCR requires a control gene which is constitutively expressed, to act both as an internal control for the PCR reaction and to allow comparisons to be made between RNA samples from different stages of *C.elegans* development. The *ama-1* gene, which encodes the large subunit of RNA polymerase II, was chosen as such a control. The mRNA accumulation of this gene is believed to be relatively constant during the larval and adult stages of development (D.Riddle, pers. comm.), and s-q rtPCR experiments performed in our laboratory with a number of cuticular collagen genes, including *dpy-7*, *sqt-1* and *col-12*, using *ama-1* as a control support this assumption (I.L.Johnstone, pers. comm.).

The amplification step of s-q rtPCR makes this method much more sensitive than Northern blot analysis. Therefore, sufficient RNA for analysis can be isolated at much more frequent intervals from small, highly synchronous populations of *C.elegans*. The ability to sample at a higher frequency than is feasible for Northern blot analysis reduces the possibility of overlooking highly restricted temporal patterns of expression of a test gene. For example, the L_1 stage lasts approximately 11 hours if worms are incubated at 25°C. During this time the worms are continuously developing, thus isolation of a single RNA sample covering the L_1 phase from a highly synchronous population will not be representative of all stages of L_1 development because different stages will be enriched depending on the time at which the RNA was extracted after hatch. Indeed, experience in our laboratory using Northern blot analysis to study the expression of cuticular collagen genes has demonstrated that expression of these genes can be missed when using wide sample windows in conjunction with highly synchronous cultures of *C.elegans*.

This problem can be reduced dramatically by increasing the frequency with which RNA samples are taken, so that each sample represents a much narrower window of time during development. For example, RNA extracted at two hour intervals after hatch would generate at least 5 samples covering the L_1 phase of development. Together, these samples will be more representative of the period of *C.elegans* development which occurs during this phase and will increase the probability of detecting transcripts from genes with highly restricted temporal patterns of expression. Furthermore, even if such RNA samples are pooled to generate samples representing each of the larval and adult phases of development, each pooled sample will be more representative of the developmental phase it covers than a single RNA sample isolated during the same phase, because there will be less enrichment for any one period of development within that phase. These factors make s-q rtPCR a powerful tool for the analysis of temporal patterns of gene expression. The approach was therefore chosen because of the reasons discussed above and because the approach had already been used very successfully in our laboratory to analyse the temporal patterns of expression of a number of cuticular collagen genes.

5.2.2. Results of Northern blot analysis

Total RNA was prepared from embryos (the embryos were not staged and therefore represent mixed stage embryos) or mixed stage N2 *C.elegans* populations (containing RNA from embryos, larvae and adults). Four aliquots of 4µg of mixed stage or embryo total RNA were electrophoresed through a 1% agarose gel under denaturing conditions and blotted to Amersham Hybond N. The filter was subsequently cut into four strips, each carrying a sample of mixed staged and embryo total RNA. The Northern blots were hybridised overnight with the cDNA clones of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. The filters were initially washed at medium stringency and subsequently at high stringency. The autoradiographic results obtained after medium stringency and high stringency washes are shown in Figures 5.1 and 5.2 respectively. Previous results, from Southern analysis with *C.elegans* genomic DNA, using the same cDNA clones as probes (Chapter 3, Section 3.2.5), demonstrated that wash stringencies similar to the high stringency wash of the Northern blots were sufficient to prevent significant cross hybridisation of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to one another or to other sequences in the *C.elegans* genome. Thus, the results obtained from the Northern blots after high stringency washing are unlikely to be a result of cross-hybridisation to additional, related sequences. The hybridisation pattern obtained after medium stringency washing (Figure 5.1) is virtually identical to that obtained after high stringency washing (Figure 5.2), the only difference being the much reduced signal intensity obtained after high stringency washing. This was probably a result of a general decrease in the amount of probe hybridising to target transcript caused by the high stringency washes (and not as a result of reduced-cross hybridisation to other, similar transcripts from related genes) since the actual pattern of hybridising bands remains unchanged. Two features of the hybridisation patterns obtained support this conclusion. First, though transcripts from *cpr-4*, *cpr-5* and *cpr-6* were detected after high stringency washing, no transcript was detected for *cpr-3* with medium stringency washing, suggesting that medium stringency washing is at least sufficient to prevent cross hybridisation of a *cpr-3* probe to transcripts from the other 3 genes. Second, though the *cpr-6* transcript is noticeably larger than both the *cpr-4* and *cpr-5* transcripts, none of the Northern blots produce two hybridising bands, indicating that cross-hybridisation of the *cpr-6* probe to *cpr-4* and *cpr-5* transcripts (or vice versa) is not occurring. It seems extremely unlikely that a cDNA probe, for example the *cpr-6* cDNA, would cross hybridise to another transcript of very similar size to that of *cpr-6* without also cross-hybridising to the transcripts of either *cpr-3*, *cpr-4* or *cpr-5*. Furthermore, the results obtained from Southern blot analysis of *C.elegans* genomic

DNA using the four cDNA clones as probes (Chapter 3, Section 3.2.2) did not reveal any indication of the presence of additional cathepsin B-like genes with extensive homology to *cpr-3*, *cpr-4*, *cpr-5* or *cpr-6* within the genome of *C.elegans*. Accordingly, I will principally discuss the results obtained after medium stringency washing

Ethidium bromide staining of the RNA gel (Figure 5.3A) prior to transfer to the Hybond N filter indicated approximately equal loading of mixed stage (lane 1) and embryo (lane 2) total RNA. Transfer of the RNA to the Nylon filter was also tested by ethidium bromide staining of the agarose gel after blotting, which revealed no visible fluorescence. The loading and transfer of RNA was also confirmed for two of the four filters by probing with a 7 kb *Bam*HI rDNA fragment containing a single repeat of the 18s and 26s ribosomal DNA sequences, gel purified from the pCe7 clone (Files and Hirsh, 1981). The results from this (Figure 5.3B) indicate that there was slightly less mixed stage RNA than embryo RNA present on the filters and that both the mixed stage and embryo RNA was degraded. However, it also demonstrated that the total RNA had been successfully transferred to the Nylon filter in these two cases, and therefore that the transfer of RNA to the other two strips was probably also successful.

Single hybridising bands, corresponding to a transcript size of approximately 1.3 kb for *cpr-4* and *cpr-5* and of approximately 1.45 kb for *cpr-6* are visible on the autoradiographs (Figures 5.1B, 5.1C and 5.1D), indicating that these three genes each generate predominant transcripts of a single size. This is consistent with the results obtained from the 5′ RACE experiments (Chapter 4, Section 4.2.6), which did not generate any evidence for transcription initiation occurring at more than one promoter region for each gene. The sizes of the mature transcripts of *cpr-4*, *cpr-5* and *cpr-6* (predicted from DNA sequencing of their cDNA and 5′ RACE clones) are 1,120 bases, 1,152 bases and 1,260 bases, respectively. As expected, these sizes are consistently smaller than the sizes obtained from Northern blot analysis since poly(A) tails were not included in the mature transcript sizes predicted by DNA sequencing. Furthermore, Northern blot analysis indicates that the mature transcript of *cpr-6* is approximately 150 bases larger than the mature transcripts of *cpr-4* and *cpr-5*, which is also in agreement with the predicted mature transcript sizes of these three genes determined by sequencing their cDNA and 5′RACE clones.

Hybridising bands corresponding to the transcripts of *cpr-4* and *cpr-5* were only visible with mixed stage total RNA (Figures 5.1B and 5.1C, lane 1). This suggests that induction of expression of these two genes occurs at some point around or after hatching of the embryo. A hybridising band corresponding to the *cpr-6* transcript is visible in both embryo and mixed stage RNA at medium stringency (Figure 5.1D). However, the signal intensity of the transcript in mixed stage RNA is much greater than that in the embryo RNA, suggesting that the abundance of the *cpr-6* transcript is much greater in mixed stage RNA than in embryo RNA. It should be noted that, by the nature of the population of *C.elegans* from which the mixed stage RNA was isolated, some embryo RNA was also present. However, since the *cpr-6* transcript abundance seen with embryo RNA is so low, the contribution made by embryo RNA to the signal seen in the mixed stage RNA must be minimal. Therefore, the increased *cpr-6* transcript abundance seen with the mixed stage RNA must be due to an increase in expression occurring at some time around or after hatch.

Since no transcript was detectable for *cpr-3* from total RNA (Figure 5.1A), poly(A)$^+$ RNA was selected from embryo and mixed stage total RNA. There was insufficient poly(A)$^+$ selected RNA to calculate the concentration by spectrophotometry and therefore all the poly(A)$^+$ RNA obtained was electrophoresed through a 1.% agarose gel and transferred to an Amersham Hybond N membrane. The poly(A)$^+$ Northern blot was hybridised overnight with a riboprobe generated from the cloned *cpr-3* cDNA (cm12b6). The Northern blot was then washed at high stringency prior to exposure to medical X-ray film. The results of this are shown in Figure 5.4. The blot was subsequently stripped and hybridised with a riboprobe generated from pCeIF to control for equal loading. This clone contains a cDNA insert encoding the *C.elegans* homologue of the eukaryotic initiation factor 4A (Roussell and Bennett, 1992) and is thought to represent one of the better controls for equal loading of poly(A)$^+$ RNA. Unfortunately, I was unable to detect any signal with this and therefore could not control for equal loading. Consequently, it was not possible to make comparisons of *cpr-3* transcript abundance between embryo and mixed stage samples.

The Northern blot analysis (Figure 5.4) reveals that *cpr-3* has a single predominant species of mature transcript. This is consistent with the 5′ RACE experiments performed with *cpr-3*, which suggest that transcription of *cpr-3* is initiated from a single promoter region upstream of this gene. The size of the mature transcript is approximately 1.6 kb. This is larger than the transcript size of 1,329 bases determined by sequencing the cDNA and 5′ RACE clones of *cpr-3* and is explained by the presence of a poly(A) tail in the mature transcript which was not included in the transcript size determined by sequencing.

In summary, Northern blot analysis revealed that the predominant transcripts of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, are each of a single size species. For each gene, the size of the mature transcript and the presence of a single size species of mature transcript is in agreement with data obtained from the 5′ RACE experiments and sequencing of the appropriate 5′ RACE and cDNA clones for each gene. For *cpr-4*, *cpr-5* and *cpr-6*, the Northern blot data also suggest that induction of expression of these genes occurs around, or after, the time of embryo hatch, since transcript abundance is greatly increased in the mixed stage total RNA sample. Such data could not be determined for *cpr-3* because controls for equal loading failed to generate hybridising signals.

5.2.3. Results from sq-rtPCR analysis

5.2.3.1. Initial analysis of the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* using sq-rtPCR

Initially, the s-q rtPCR approach outlined in Section 5.2.1 was used to analyse the relative transcript abundance of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* in the 20 staged cDNA samples generated by Dr I..L.Johnstone (see Chapter 2, Section 2.13). These samples were generated from total RNA isolated at two hour intervals during the four larval and early adult stages of *C.elegans* development. For each gene, 23 cycles of PCR were performed using cDNA template from each of the 20 samples and primer pairs which produced recognisable size differences between the PCR products generated from cDNA and contaminating genomic DNA. This size difference was achieved by selecting primers specific to coding sequence either side of introns. The nucleotide sequences and the positions of each primer pair are given in Chapter 2, Section 2.13.1. The entire PCR reaction for each time point was electrophoresed through an agarose gel, Southern blotted to Hybond N (Amersham International plc) and probed using the same primers as used for the PCR amplification step (5′ end-labelled using polynucleotide kinase). The filters were washed at medium stringency prior to exposure to medical X-ray film. The levels of radioactivity were not assayed by scintillation and accordingly all steps outlined in Section 5.2.1 subsequent to the generation of each autoradiograph were omitted.

Only 23 cycles of PCR were used to ensure that the PCR reactions were stopped whilst still in the log phase of amplification. This would reduce the probability of misrepresentation of relative abundance caused by premature depletion of primer pairs in the PCR reactions. For each gene, 23 cycles of PCR yielded very low amounts of each product, with the products being either not or barely visible on ethidium bromide stained agarose gels prior to Southern blotting. In contrast, 35 cycles of PCR using a mix of all 20 cDNA samples and the primer pair for *cpr-3* generates clearly visible bands (data not shown). This suggests that the amplification obtained after 23 cycles of PCR is sufficiently low such that the reactants are still in excess post amplification and are therefore not rate limiting.

One concern with the amplification step of s-q rtPCR is that excessive amplification of inappropriate template may affect the quantitative nature of this approach, by acting as a sink for the PCR primers and so reducing their effective concentration in the reaction mixture. Therefore, the same oligonucleotides used for the PCR amplification steps were also used as probes for hybridisation to the Southern blots of the PCR amplified products. This approach allows all the more abundant PCR amplification products, whether derived from appropriate or inappropriate template, to be visualised. The results obtained from this Southern blot analysis (Figures 5.5, 5.6, 5.7 and 5.8) reveal that all the predominant hybridising bands are of the expected size for amplification from the cDNA and genomic DNA templates of the appropriate target gene and of *ama-1*. This suggests that each primer pair is capable of specifically amplifying the appropriate template. For the same reason, these data suggest that none of the target gene primer pairs are incompatible with the *ama-1* primer pair, since none of these combinations behaved specifically in combined reactions (i.e. they did not produce additional predominant products other than those expected). However, the *cpr-4* primer pair did produce several weakly hybridising bands of inappropriate sizes. Though none of these bands individually generated strong hybridising signals, the number of bands produced suggested that amplification of inappropriate templates could potentially act as a sink for the *cpr-4* primer pair.

The autoradiographic data (Figures 5.5, 5.6, 5.7 and 5.8) reveal several important features regarding the relative transcript abundance of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* during larval and adult stages. By comparing the signal intensity of the band corresponding to the transcript of the target gene to that of the *ama-1* internal control, the relative transcript abundance of the target gene in each sample can be assessed. Such comparison at each time point for each gene reveals that there is no dramatic increase in expression occurring for any of the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, between any two adjacent time points during the larval and adult stages of *C.elegans* development. In contrast, the autoradiographic data obtained for the collagen gene *dpy-7*, using the same approach (Figure 5.9, from I.L.Johnstone), reveals a large increase in *dpy-7* transcript levels, with respect to *ama-1*, which recurs at a similar time point prior to moulting at each larval stage. Therefore it appears that none of the four genes, *cpr-3*, *cpr-4*, *cpr-5* or *cpr-6*, show a rapid induction of expression at any point within the series, nor do they show any evidence of cyclical patterns of expression during *C.elegans* development. Interestingly, the signal intensity of the band corresponding to the *cpr-3* transcript shows less variation through the larval and adult stages of *C.elegans* development than *ama-1*. This suggests that *cpr-3* transcript abundance may fluctuate less during *C.elegans* larval and adult development than *ama-1*.

In summary, the initial s-q rtPCR experiments indicate that the parameters chosen are appropriate for analysis of the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. The results generate no evidence of highly restricted temporal expression patterns for the four genes, with no rapid and short-lived increases in relative transcript abundance occurring within any one developmental stage. The results also show no evidence of a cyclical pattern of expression during the larval and adult stages of *C.elegans* development. Such patterns have been observed for several cuticular collagen genes, including *dpy-7* (Figure 5.9) which imply induction at one point within each larval stage. Since cuticular collagens are required for synthesis of the new cuticle prior to each larval moult, the periodic expression patterns observed with those cuticular collagen genes analysed, suggest that each larval moult may follow a similar programmed sequence of events. The cathepsin B enzyme has been shown to be capable of degrading collagen and is thought to perform a role in bone remodelling in vertebrates (Chapter 1, Section 1.7.4). Accordingly, it may have been possible that the cathepsin B-like gene family evolved in nematode species to perform an analogous role in remodelling the cuticle. However, the absence of such a cyclical pattern of expression for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* during the larval stages of development suggest that this is not the case.

5.2.3.2. Adaptations made to the initial s-q rtPCR protocol

From the initial s-q rtPCR experiments, it is apparent that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* do not show the type of expression pattern observed with *dpy-7*. With none of the genes showing rapid induction of expression, highly temporally restricted expression within any one developmental stage or cyclical expression. However, it is possible that these genes might possess more subtle temporal expression patterns during *C.elegans* development, with gradual changes in the relative transcript abundance occurring over several samples. Accordingly, the s-q rtPCR protocol was adapted to enable experiments to be performed which could determine if such gradual fluctuations were occurring.

In order to analyse subtle changes in transcript abundance between stages, it is essential that real differences in transcript abundance can be distinguished from differences arising from experimental error. This is especially important with s-q rtPCR because the introduction of a PCR amplification step makes this protocol inherently more complex than approaches such as Northern blot analysis, thereby increasing the probability of introducing error. Accordingly, the s-q rtPCR protocol was adapted to address this problem and the adaptations made are discussed below.

The s-q rtPCR approach was simplified to make the procedure more reliable. Accordingly, the number of cDNA samples analysed in each experiment was reduced. The initial s-q rtPCR experiments indicate that none of the four genes exhibited highly restricted patterns of temporal expression within any one developmental stage. Therefore, the cDNA samples covering each stage of development were pooled to generate five amalgamated samples each representing the L_1, L_2, L_3, L_4 or adult stage, respectively. An equal volume of cDNA from each time point was amalgamated as follows: samples 1-6 for L_1, samples 7-9 for L_2, samples 10-12 for L_3, samples 13-15 for

L_4 and samples 16-20 for early adult. Equal volumes of each appropriate cDNA sample were used to generate the amalgamated samples because the results from the initial s-q rtPCR experiments (Figures 5.5, 5.6, 5.7 and 5.8) suggest that no one sample contains significantly higher concentrations of cDNA than any other. First, the fluctuations in signal intensity corresponding to *ama-1* cDNA template between different samples are not consistent among the four genes, suggesting that such fluctuations do not reflect differences in the cDNA concentration in each sample, but rather differences in the efficiency of amplification. Second, the signal intensities corresponding to *cpr-3* cDNA template remain remarkably constant between the 20 samples (Figure 5.5), suggesting that each of the cDNA samples from different time points contain similar amounts of *cpr-3* template. Such similar levels of *cpr-3* template in each sample suggest that the transcript abundance of this gene remains relatively constant during the larval and adult stages of *C.elegans* development. Accordingly, this suggests that each sample contains similar quantities of cDNA.

A new primer pair was designed for *cpr-4* (Chapter 2, Section 2.13.1), because the previous primer pair generated some non-specific amplification products. Since *cpr-4* possesses only a single, small (49 bp) intron, the small size difference between the amplification products derived from cDNA template and contaminating genomic DNA would make it difficult to separate these products by agarose gel electrophoresis. Accordingly, the primer pair was designed to only amplify cDNA template, by designing one of the primers to span the intron site within the cDNA. The new *cpr-4* primer pair was tested for its ability to specifically amplify *cpr-4* cDNA template in the presence of *ama-1* primers using 35 cycles of PCR, mixed stage cDNA and the same annealing temperature to be used in the s-q rtPCR experiments. The PCR reaction generated three PCR products. The sizes of these products were as expected for amplification of the *cpr-4* cDNA template by the *cpr-4* primer pair and amplification of the *ama-1* cDNA template and contaminating genomic DNA by the *ama-1* primer pair, indicating that this new primer pair was capable of specifically amplifying *cpr-4* cDNA in the presence of *ama-1* primers.

The hybridisation protocol of s-q rtPCR was altered to allow higher stringency washes because previous strategies using 5′ end-labelled oligonucleotide probes produced filters with background signals that were too high and too variable across the surface of the filter to allow accurate analysis by scintillation counting. Therefore, probes for *cpr-3*, *cpr-4*, *cpr-5*, *cpr-6* and *ama-1* were generated by PCR amplification (Chapter 2, Section 2.13.3) using the same primer pairs to be used in the s-q rtPCR experiments. The larger size of these probes allowed the use of high stringency washes which reduced the levels and variability of background radioactivity after hybridisation of the probes to the Southern blots of the PCR amplified products.

A control for efficient transfer of the PCR amplified DNA to the Hybond N (Amersham International plc) filters during Southern blotting was introduced to ensure that variation in the relative transcript abundance between samples was not a result in variation in the efficiency of transfer of the PCR products to the filters. Therefore, the size of each PCR reaction was doubled (to a 50µl final volume). After the amplification step, the reaction from each time point was split into two 20µl aliquots to generate two sets of samples from the same PCR reaction. Each set of samples was electrophoresed through an agarose gel and transferred to Hybond-N (Amersham International plc) by Southern blotting. The filters were then probed and washed in identical conditions prior to exposure to medical X-ray film. By comparing the autoradiographs generated by the two blots derived from the same PCR reaction, it was possible to assess the efficiency of transfer since each blot should give virtually identical results.

An additional cDNA sample covering the embryo stages of *C.elegans* development was introduced because Northern blot analysis suggests that induction of expression of *cpr-4*, *cpr-5* and *cpr-6* occurs around or after the time of embryo hatch and the 20 staged cDNA samples used in the initial s-q rtPCR experiments only covered the larval and adult stages of development. First strand cDNA was generated from 2µg of mixed stage embryo RNA (Chapter 2, Section 2.13). The cDNA was then diluted to an appropriate concentration for use in s-q rtPCR. This dilution was determined using 23 cycles of PCR with primer pairs for *ama-1* and *cpr-3* and 10 fold serial dilutions of the mixed stage embryo cDNA. The PCR products were electrophoresed through an agarose gel and visualised by ethidium bromide staining. This identified faint PCR products generated from the 100 fold dilution and therefore a 500 fold dilution was used in the s-q rtPCR experiments. It is important to note that there is no currently available method for synchronising *C.elegans* embryos. It is therefore not possible to determine what proportion of each embryo stage is represented in the embryo cDNA sample. This sample is therefore distinct from the amalgamated samples in this respect. However, the RNA was extracted immediately after isolation of the embryos and therefore is unlikely to contain RNA from L_1 larvae, since only a very small fraction of the embryos, at the end of embryo development, will have had time to progress to the L_1 stage.

Finally, for each gene, the s-q rtPCR experiments were performed in triplicate to assess the reproducibility of each set of s-q rtPCR experiments. This allowed the use of statistical tests to help determine whether differences in relative transcript abundance between samples reflected genuine differences or merely experimental error.

5.2.3.3. Summary of the modified s-q rtPCR approach used

There follows a brief summary of the experimental procedure used to obtain the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* or *cpr-6* (the procedure is also outlined in Figure 5.10). Three independent sets of PCR reactions (A, B and C) were performed with primers specific for the target gene and for *ama-1* using the mixed stage embryo cDNA and the amalgamated cDNA samples for L_1, L_2, L_3, L_4 and adult stages. Each set of PCR reactions was split into two 20 µl aliquots and electrophoresed through an agarose gel. The agarose gel was then cut into two, with each half containing amplification products from each of the six developmental stages, derived from the same set of PCR reactions. Each half of the gel was then Southern blotted to Hybond N filters (Amersham International plc) to give a total of six filters carrying the amplification products from three independent PCR reactions. Therefore, for each gene, reaction A yields filters 1 and 2, reaction B yields filters 3 and 4 and reaction C yields filters 5 and 6. The six filters for each gene were simultaneously hybridised with probes specific to the target gene and to *ama-1*. The filters were washed at high stringency and exposed overnight with autoradiographic film. The resulting six autoradiographs were used as a template to excise the hybridising bands from the filters which corresponded to the target gene cDNA and *ama-1* cDNA amplification products. For each sample, the radioactivity of the two bands was measured by counting in scintillant. The relative transcript abundance of the target gene was expressed as the ratio of the counts per minute (cpm) obtained from the target gene cDNA amplification products to the cpm obtained from the *ama-1* cDNA amplification products.

5.2.3.4. The modified s-q rtPCR approach generates reproducible results

The autoradiographic data obtained from the six filters for each of the four genes are shown in Figures 5.11, 5.12, 5.13 and 5.14. The data obtained from scintillation counting, including the relative transcript abundance ratios, for each of the four genes are shown in Tables 5.1, 5.2, 5.3 and 5.4. The relative abundance ratios obtained are summarised graphically in Figure 5.15.

For each gene, each pair of filters, from a common PCR reaction, shows a very similar pattern of hybridisation on the autoradiographs (Figures 5.11, 5.12, 5.13 and 5.14), including the relative intensity of signals between different bands on the same filter. This similarity is also reflected in the graphical representations of the relative transcript abundance ratios (Figure 5.15) where each pair of graphs shows similar trends. This indicates that the PCR products from each reaction were transferred to the Nylon filters in a reproducible manner during Southern blotting. Therefore, the variation in transcript abundance of each gene during *C.elegans* development, discussed later, is not a result of variations in transfer of the PCR products. While each pair of autoradiographs shows a very similar pattern of hybridisation, the overall signal intensity may differ substantially. This cannot be easily explained since each pair of filters was hybridised with the same probe in a single hybridisation bottle. This phenomenon does not appear to affect the results since this variation occurs to a similar extent across the entire filter.

To assess the reproducibility of the amplification step of s-q rtPCR, three independent pools of cDNA from each developmental stage were amplified in different PCR reactions. Again, looking at the autoradiographic results (Figures 5.11, 5.12, 5.13 and 5.14), it is apparent that each filter pair shows a very similar pattern of hybridisation to the other filter pairs obtained from different PCR reactions for the same gene. The three sets of s-q rtPCR reactions were capable of generating not only a similar pattern of bands after hybridisation but also a similar pattern in the relative intensities of these bands, with respect to other bands on the same autoradiograph. This reproducibility is also reflected in the graphical representations of the relative transcript abundance ratios (Figure 5.15). Each graph contains three data sets, obtained from three independent

PCR reactions (A, B and C) for each gene. Each data set produces a similar trend in the temporal pattern of expression for a given gene and therefore indicates that the PCR amplification step is reproducible, as suggested by the autoradiographic data. The scintillation samples from filters 3 and 5 for *cpr-6* were lost and therefore graphical comparison of data obtained from these filters was not possible. However, the original autoradiographic data (Figure 5.14) clearly demonstrate that the three independent PCR reactions for this gene generated very similar results, suggesting that the PCR amplification step for *cpr-6* is also reproducible. These data provide strong evidence that the PCR amplification step of s-q rtPCR generates reproducible results for each gene.

5.2.3.5. The temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* obtained from the modified s-q rtPCR approach

As discussed above, the controls implemented suggest that the transfer of DNA to Hybond N (Amersham International plc) with the modified s-q rtPCR approach is reliable. Therefore, one set of triplicate results (obtained from the three independent PCR reactions, A, B and C) was used to analyse the temporal expression pattern of each gene. Thus, the scintillation data obtained from filters 1, 3 and 5 were analysed for each of the three genes, *cpr-3*, *cpr-4* or *cpr-5*. For *cpr-6*, there was no scintillation data for filters 3 and 5 and therefore filters 2, 4 and 6 were analysed. The mean and standard deviation of the relative transcript abundance ratio was determined for each developmental stage, from the three independent PCR reactions performed for each gene (Tables 5.1, 5.2, 5.3 and 5.4). These results are shown graphically in Figure 5.16 and reveal some very interesting differences between the temporal expression patterns of each gene.

The results suggest that all four genes show an induction of expression at the time of hatch but this induction varies in size between the four genes. Subsequent to this induction, the genes show quite different temporal patterns of expression. Thus, maximal expression of *cpr-3* occurs at L_1, while *cpr-4* shows two peaks at L_1 and L_4, *cpr-5* shows one major peak at L_2 and a smaller peak at L_4, and *cpr-6* shows one peak in the adult. I used Student's *t*-test to determine if these peaks of expression were significant. This test uses, as its null hypothesis, the assumption that two samples have been derived from the same population, and therefore that any differences in the mean and standard deviation of these samples can be accounted for by sampling errors. Using this test, I compared the relative transcript abundance of each stage with the next stage in development (i.e. embryo with L_1, L_1 with L_2, L_2 with L_3, L_3 with L_4, L_4 with adult). The results of this are shown in Table 5.5. The *t*-test confirms most of the observations made by eye. For *cpr-3*, the most obvious variation is the increase in transcript abundance between the embryo and L_1 stages, which is also found to be significant using the *t*-test (Table 5.5). None of the other fluctuations were found to be significant, supporting the initial s-q rtPCR data (Figure 5.5) which suggest that *cpr-3* transcript abundance remains fairly constant through the larval and adult stages of *C.elegans* development. The temporal expression pattern for *cpr-4* is very similar to that of *cpr-3* with the exception of a marked decrease in relative transcript abundance between the L_4 and adult stages. These data suggest that *cpr-4* is expressed predominantly in the larval stages of *C.elegans* development. The *t*-test supports this conclusion, since it identifies the increase in abundance between the embryo and L_1 stages and the decrease in abundance between the L_4 and adult stages as being highly significant and significant respectively (Table 5.5), but does not find any other variations to be significant. For *cpr-5*, there is a dramatic (11.7 fold) increase in relative transcript abundance between the embryo and L_1 stages and the relative transcript abundance continues to rise to a maximum in the L_2 stage. There is a general downward trend (interrupted by a small peak at L_4) in transcript abundance for *cpr-5* in the subsequent larval and adult stages. The *t*-test found the 11.7 fold increase in relative transcript abundance between the embryo and L_1 stages to be significant. However, it also found all the differences in relative transcript abundance between subsequent stages to be significant. This is surprising because there is only a maximum 2.3 fold variation in relative transcript abundance between these subsequent stages. *cpr-6* shows a very different temporal pattern of expression to the other three genes. There is a slight increase in relative transcript abundance between the embryo and L_1 stages however this is overshadowed by a much greater increase between the L_3 and L_4 stages which continues to rise to a maximum in the early adult. This was confirmed using the *t*-test which did not find the increase between the embryo and L_1 stages as significant but identified the increase between L_3 and L_4 stages as being highly significant. The temporal expression profile for *cpr-6* therefore suggests that this gene is not expressed at significant levels until around the L_3 stage, unlike the other three genes where induction appears to occur around the time of embryo hatch.

There are not only variations in the timing of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* but also in the degree of induction that occurs. It should be noted, before discussing these variations, that the scales of the Y-axes of the graphs shown in Figure 5.16 vary to allow comparison of the four genes. The four genes can be divided into two groups on the basis of their degree of induction. *cpr-3* and *cpr-4* show only 2.3 and 5.3 fold increases between minimal and maximal expression while *cpr-5* and *cpr-6* show 27.0 and 17.9 fold increases, respectively. Therefore, for *cpr-3*, the differences in the relative transcript abundance seen during *C.elegans* development may only represent slight variation from a gene that is essentially constitutively expressed. In contrast, though *cpr-4* does not show great variation between minimal and maximal transcript abundance, the data here suggest that this gene is not constitutively expressed, but upregulated during the larval stages of development. In contrast, *cpr-5* and *cpr-6* are highly developmentally regulated. *cpr-5* appears to be induced around hatch, and its relative transcript abundance increases dramatically up to the L_2 stage, where it begins to decrease. *cpr-6* shows quite a different pattern, with a dramatic induction of expression occurring at the late L_3 / L_4 stage, and relative transcript abundance increasing into the early adult stage.

It is not possible to quantify differences in absolute transcript abundance between genes using a s-q rtPCR approach since there are many factors which can influence the relative efficiency of amplification of the target transcript with respect to the *ama-1* transcript. In particular, the efficiency with which primer pairs can amplify their target sequence has a significant effect on the perceived abundance of that sequence with respect to *ama-1*. However, it is interesting to note that *cpr-5* and *cpr-6*, which both show the highest degrees of induction, also have the highest relative transcript abundance ratios with respect to *ama-1*. At the developmental stage where maximal transcript abundance is seen for *cpr-5* and *cpr-6*, they show a 7.6 and 3.1 fold greater abundance over *ama-1*. The correlation between greater degrees of induction and levels of expression with respect to *ama-1* does suggest that *cpr-5* and *cpr-6* transcripts may be more abundant than those of *cpr-3* and *cpr-4*, once induction has occurred.

5.2.4 Comparison of the results obtained from Northern and s-q rtPCR analysis

It is essential that a constitutively expressed gene is used as the internal control for s-q rtPCR. Though *ama-1* mRNA abundance appears to remain relatively constant during the larval and adult stages of *C.elegans* development, recent reports suggest that *ama-1* transcripts may be more abundant in the embryo (D.Riddle, pers. comm.). If *ama-1* levels increased, or decreased, dramatically during any of the stages of *C.elegans* development one would expect to see a conserved dip, or peak, in the relative transcript abundance as measured by this method occurring at the same stage for each of the four genes. No such conserved dip, or peak, in expression levels is seen for these genes during the larval and adult stages indicating that *ama-1* levels remain relatively constant during these stages. However, all four genes show an increase in expression between the embryo and L_1 stages. This could represent a real induction event occurring around the time of hatch, or an artefact caused by increased *ama-1* transcript levels in the embryo, giving rise to an artefactual perceived increase at hatch. The increase in relative transcript abundance between the embryo and L_1 stages for *cpr-4* and *cpr-5* identified by s-q rtPCR is confirmed by the results obtained from Northern blot analysis. Northern blot analysisfor *cpr-4* and *cpr-5* (Figures 5.1B and 5.1C) detected no evidence of transcripts for these two genes in the embryo. In contrast, transcripts for both genes were detected in mixed stage RNA. The s-q rtPCR analysis detected transcripts for both these genes in all stages of *C.elegans* post hatch. For *cpr-4*, this expression was predominantly in the larval stages while , for *cpr-5*, high levels of expression were seen in both the larval and adult stages. Together, the data suggest that both these genes are expressed at very low levels in embryos and at much higher levels in the L_1 stage. This provides strong evidence that induction of expression for these genes is occurring at some point around the time of hatch.

For *cpr-6*, transcripts were detected in both embryo and mixed stage RNA by Northern blot analysis (Figure 5.1D). However, the Northern blot data suggested that the transcript was in far greater abundance in the mixed stage RNA. The s-q rtPCR experiments for *cpr-6* generate two peaks of expression, one small peak at L_1 and a much larger peak starting between the L_3 and L_4 stages and rising in the early adult stages. Statistical analysis using the *t*-test suggest that only the large peak starting between the L_3 and L_4 stages is significant. Together, the Northern blot and s-q rtPCR data suggest that *cpr-6* is transcribed at a very low level through the embryo and early larval stages and then induced at some point around the L_3 stage.

For *cpr-3*, the Northern blot results could not be used to obtain any information regarding variation in transcript abundance between embryo and mixed stage RNA samples because the control for equal loading failed to generate a signal. The s-q rtPCR results suggest that this gene is expressed constitutively during the larval and adult stages of *C.elegans* development. In addition, there is only a 2.3 fold increase in relative transcript abundance between the embryo and L_1 stages. These data alone suggest that *cpr-3* is essentially constitutively expressed during all stages of *C.elegans* development, although there may be a slight increase in expression between the embryo and L_1 stages.

In summary, the Northern blot and s-q rtPCR data indicate that each of the four genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* show different patterns of temporal expression during *C.elegans* development. *cpr-3* is expressed at all stages of development and shows only a slight increase of expression at hatch. *cpr-4* appears to be induced around hatch and is expressed predominantly in the larval stages. *cpr-5* and *cpr-6* both show highly developmentally regulated patterns of expression. Both of these genes show over 15 fold increases in their transcript abundance between minimal and maximal expression but the timings of these inductions are quite different. *cpr-5* is induced around the time of hatch, and shows relatively high levels of expression in the subsequent larval and adult stages. *cpr-6* appears to be transcribed at a low level for most stages of *C.elegans* development but is subsequently induced around the L_3 stage, with expression levels continuing to rise through the L_4 and early adult stages. The results reported here indicate that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encode enzymes which are expressed during different stages of *C.elegans* development. Two possible situations best explain the variation in the temporal expression patterns observed between the four genes. In one situation, these genes might encode cathepsin B-like enzymes which perform distinct biological processes in different tissues, during different stages of *C.elegans* development. In the other situation, the enzyme products may be expressed in the same tissues and may perform the same or similar biological functions, but are required in different combinations during *C.elegans* development. The diverged predicted amino acid sequences of the cathepsin B-like enzymes encoded by the four genes (Chapter 4) are in agreement with either of these situations, since in both cases one would expect the enzymes to have different activities and substrate specificities, either to degrade / process different substrates in different tissues or to degrade / process different substrates (or combinations of substrates) in the same tissue but at different stages during *C.elegans* development.

5.3. Analysing the spatial expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

5.3.1. Overview of the *lacZ* fusion transgene approach

lacZ fusion transgenes were used to analyse the spatial expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. In brief, the 5′ flank and part of the coding region of each of the four genes was ligated upstream of, and in-frame with, the bacterial *lacZ* ß-galactosidase gene, to produce translational gene fusions. Plasmid constructs containing these gene fusions, along with a marker plasmid, pRF4, were microinjected into the syncytial gonad of adult hermaphrodite worms. Single F2 progeny which transmitted the marker plasmid were subsequently selected and transgenic lines generated from these. The method used (Mello *et al.,* 1991) generates a large extrachromosomal array made up mostly of tandem repeats of the construct and the pRF4 marker plasmid. Analysis of such arrays has demonstrated that they vary in size, being made up of a total of approximately 80 - 300 copies of the injected plasmids (Stinchcomb *et al.,* 1985). The arrays are not inherited in a Mendelian manner because they are not attached to chromosomes. However, transgenic lines can be kept indefinitely by selecting for the mutant phenotype produced by the marker plasmid..

5.3.2. Advantages of using *lacZ* transgene fusions

The *lacZ* transgene fusion approach is the method of choice for initial analysis of the spatial expression patterns of genes in *C.elegans*. The popularity of this method is based on several reasons. First, meaningful results are obtained with a high frequency. Second, the approach allows analysis of the regulatory regions responsible for controlling expression of the genes being studied. Third, a large number of genes have been tested using *lacZ* transgene fusions resulting in this approach being well characterised. Finally, extensive use of this method has revealed that it is relatively free of serious artefacts.

For a number of genes, the spatial expression pattern generated by *lacZ* transgene fusions has also been compared to the spatial expression pattern of the endogenous gene by using other approaches such as *in-situ* hybridisation or antibody staining. In many cases the spatial expression patterns produced by the gene fusions accurately reflect the spatial expression patterns of their endogenous genes. For example, *lacZ* fusions with the MyoD family homologue *hlh-1* faithfully reproduce the tissue-specific expression pattern of the endogenous gene, in body wall muscle precursors (Krause *et al.,* 1994). Okkema *et al* (1993) have also demonstrated that *lacZ* gene fusions with *unc-54*, *myo-1*, *myo-2* and *myo-3* all mimic the tissue specific expression patterns of their endogenous counterparts. The *unc-54*::*lacZ* and *myo-3*::*lacZ* fusions exhibit expression in the body-wall muscle and specialised body muscles associated with the intestine, the vulva and the uterine sheath while the *myo1*::*lacZ* and *myo-2*::*lacZ* fusions exhibit expression only in pharynx muscle. Hamelin *et al* (1992) have demonstrated that a *lacZ::mec-7* ß-tubulin gene fusion is expressed specifically in the touch neurones of transgenic worms which also react with anti-*mec-7* antibodies. In other cases, the pattern of expression of test gene::*lacZ* fusions have correlated well with mutant phenotypes resulting from mutations in the endogenous gene whose expression is being analysed. For example, the cuticular collagen gene *dpy-7*, which has a dumpy mutant phenotype resulting from cuticular defects, has been shown to be expressed in the hypodermis, the site of cuticle synthesis (I.L.Johnstone, pers.comm.). More recently, Tabish *et al* (1995) have demonstrated that *lacZ* gene fusions with the *osm-3* kinesin gene exhibit expression in chemosensory neurons open to the environment. The authors have correlated this pattern of expression with the *osm-3* mutant phenotype, which shows structural defects in the amphid and phasmid sensilla. Such examples, suggest that in many cases *lacZ* gene fusions with a gene of interest accurately reflect the pattern of expression of the endogenous gene.

Other techniques are available for analysing the temporal expression patterns of genes. The two principle methods are detection of RNA *in-situ* (*in-situ* hybridisation) and antibody staining of whole, or sectioned, *C.elegans*. Whole worm *in-situ* hybridisation has been used to analyse the spatial expression patterns of genes within embryos. However, this approach has proved problematic when used to detect RNA from developmental stages other than embryos. This is thought to be a result of the cuticle acting as a barrier which prevents equal access of the nucleic acid probes to different tissues within the worm. Antibody staining is a very powerful method because it is capable of determining the spatial distribution of the final protein product of a gene, as opposed to determining only where a gene is expressed. In addition, immunogold labelling of antibodies and electron microscopy allow the subcellular distribution of target proteins to be determined. However, these approaches require the generation of specific immunological reagents. The generation of such reagents may be very time consuming and there is no way of guaranteeing that the antibodies obtained will be specific. Furthermore, this problem can be exacerbated when studying a multigene family as it may not be possible to generate antibodies specific for each member of the family.

5.3.3. Problems associated with the use of *lacZ* transgene fusions

Though *lacZ* reporter gene fusions often reliably reflect the expression patterns of endogenous copies of the genes being tested, a few problems with this approach have been observed. One of the most commonly observed problems is that of mosaic expression. This is in part due to the maintenance of the transgene as an extrachromosomal array in transgenic worms. Because these arrays are not attached to chromosomes, they may be lost during somatic cell divisions as a result of random segregation between daughter cells, and therefore may give rise to mosaic animals. Such mosaicism will cause variation in the cells which express the transgenes. However, variable staining patterns may also arise as a result of other factors such as incomplete fixation or the inability of all the cells which should stain to express the transgene at sufficient levels to be detected. Since such factors generally result in variable expression patterns similar to those generated by 'true' mosaicism, this term is often used to describe variable staining patterns with *lacZ* transgene fusions although the true basis of the variability may not be known. Mosaicism (staining variability) has been reported for a number of different gene fusions studied in *C.elegans*. For example, the P-glycoprotein genes, *pgp-1* and *pgp-3* (Lincke *et al.,* 1993) and the cuticular collagen gene *dpy-7* (J.Gilleard, pers. comm.) have both been shown to exhibit mosaic expression as *lacZ* gene fusions in the intestine and hypodermis respectively. It is important to note that valid conclusions regarding the tissue-specific expression pattern of an endogenous gene may still be made from data obtained using *lacZ* gene fusions that exhibit mosaic expression. Such conclusions can be made because mosaicism generally results in staining of different cells from the same tissue in different animals of the same transgenic line. By comparing a number of different animals from the same line, therefore, it is possible to determine the tissues that express the fusion transgene, but not particular cells within the tissue that express the fusion transgene.

In some cases, ectopic expression of *lacZ* gene fusions in additional tissues outside those normally expressing the endogenous gene has also been reported. For example, ectopic expression of *hlh-1* (which is expressed in the mature body wall muscle) has been observed in the pharynx, posterior larval gut and in embryonic hypodermal precursors, when using very short promotor sequences and *lacZ* transgene fusions (Krause *et al.,* 1994). Such ectopic expression may result from the absence of all the regulatory sequences necessary for appropriate expression of the transgene fusion when using very short 5′ flank sequences. Alternatively, it is thought that such ectopic expression may also arise from interaction between sequences in the plasmid backbone and those in the promotor region being analysed when in close proximity to one another.

Another problem to be considered when using *lacZ* gene fusions is the apparent inability of such fusions to exhibit expression in the embryonic, larval or adult germ line of *C.elegans*. The analysis of expression of the small heat shock (*hsp16*) genes using *hsp16*::*lacZ* gene fusions has revealed that *lacZ* fusions can be expressed in virtually all tissues of the worm, including cells of the gut, nervous system hypodermis and pharynx, however in no instance has germ line expression been demonstrated (Fire *et al.,* 1990; Stringham *et al.,* 1992). Furthermore, other transgenes, such as *msp-1* which encodes a sperm specific product, have also failed to be expressed in this nematode species.

Two other problems are associated with the use of *lacZ* transgene fusions: first, no transgene expression has been demonstrated in pre-twelve-cell embryos of *C.elegans* to date; second, expression in a single tissue or population of cells may not be seen with the transgene, though the endogenous gene is known to be expressed in that region (A.Fire, pers. comm.).

The problems discussed above have lead to the very recent introduction of a new set of modified *lacZ* reporter vectors (A.Fire, pers. comm.). These vectors essentially differ from the old vectors by the addition of one to twelve introns into the *lacZ* coding region. The addition of these introns can stimulate expression by two to three orders of magnitude making these vectors substantially more sensitive than the previous set. *In situ* localisation experiments using *unc-54*::*lacZ* constructs containing 0,1,8 and 12 intron sequences indicates that the addition of intron sequences into the *lacZ* coding region results in a large increase in the amount of *unc-54*::*lacZ* transcripts seen in the cytoplasm. This suggests that introns may increase the efficiency of transport of transcripts out of the nucleus, or may increase processing efficiency and/or transcript stability. In addition to the increased sensitivity, these new vectors also exhibit reduced levels of mosaicism, or staining variability, compared to the previous set of vectors. However, these vectors have not overcome the problems of ectopic expression or lack of expression in the germ line encountered with *lacZ* fusion transgenes. Indeed, the increase in sensitivity has also resulted in increased background expression in the pharyngeal and gut tissues, especially with vectors containing 12 introns in the *lacZ* coding region. However, the use of a decoy minigene (discussed in the following section) upstream of the 5′ Polylinker has eliminated such background expression from promoterless *lacZ* fusion vectors by preventing readthrough transcription.

5.3.4. *lacZ* fusion vectors

Both vectors used to construct the reporter gene fusions are members of a modular set of *lacZ* fusion vectors (Fire *et al.,* 1990). These vectors are well characterised because many of the studies of gene expression in *C.elegans* using *lacZ* transgene fusions have utilised them. The basic vector consists of a pUC19 derived plasmid backbone, a 5′ multiple cloning site (MCS) or polylinker and the *E.coli trps::lacZ* gene fusion. A number of cassettes can then be added to this, containing one of the following; a synthetic intron, an ATG start codon and nuclear localisation signal, a synthetic transmembrane segment and finally, a 3′ MCS, the *unc-54* 3′ end cassette or the SV40 early 3′ end cassette. I used the two *lacZ* fusion vectors, pPD21.28 and pPD95.03 (see Figure 5.17). The pPD21.28 vector comprises a pUC19 plasmid backbone, a 5′ polylinker, a synthetic intron and an SV40 nuclear localisation signal all upstream of the *E.coli trpS::lacZ* ß-galactosidase coding region. An *unc-54* 3′ end supplies the cleavage / polyadenylation signal downstream of the coding region. The pPD 95.03 vector is based on the pPD21.28 vector and therefore shares many similarities. The two important differences are the insertion of twelve synthetic introns into the *E.coli* TrpS::*lacZ* ß-galactosidase coding region. The insertion of introns into the *lacZ* gene has resulted in an increase in sensitivity by two to three orders of magnitude, making this vector significantly more sensitive than pPD21.28. The increase in sensitivity has also resulted in an increase in background, especially in the pharynx and the gut. To combat this, a decoy sequence has also been inserted just upstream of the 5′ polylinker. The decoy sequence consists of a short intron followed by a short open reading frame and a consensus *C.elegans* translational start. The open reading frame terminates just upstream of the 5′ end of the polylinker. Such a decoy sequence has been shown to prevent background expression from promoterless vectors by preventing readthrough transcription originating from the plasmid backbone (A.Fire, pers. comm.). It should be noted that the presence of a nuclear localisation signal in the two *lacZ* fusion vectors used will result in localisation of the fusion protein to the nucleus. Therefore such localisation does not reflect the actual subcellular localisation of the endogenous proteins encoded by the genes being analysed.

5.3.5. Generation of the *lacZ* gene fusions

Several factors require consideration when generating *lacZ* gene fusions. First, a sufficient amount of 5′ flanking sequence for each gene must be inserted into the *lacZ* fusion vector to ensure that all regulatory elements required for correct expression of these genes are present in the flanking sequences. Promotor analysis studies with other genes such as *dpy-7* (J.Gilleard, pers. comm.), *vit-2* (MacMorris *et al.,* 1992) and *mec-3* (Xue *et al.,* 1992) suggest that as little as 250bp - 350bp of 5′ flank sequence may be sufficient for driving appropriate spatial expression of a number of *C.elegans* genes. However, I decided to use substantially larger 5′ flank regions from *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* for initial analysis of the spatial expression patterns of these genes. The use of larger 5′ flank regions places the *lacZ* fusion transgene in a more natural environment. It increases the probability of obtaining all the necessary regulatory elements to drive correct expression of each gene. It also reduces the likelihood of ectopic expression by increasing the distance between the promoter elements in the 5′ flank and sequences in the plasmid backbone. The interaction of promoter elements and sequences in the plasmid backbone may occur when these sequences are in close proximity to one another. Such interactions are thought to cause some of the ectopic expression patterns observed when using very short 5′ flank regions (Krause *et al.,* 1994). Thus, I decided to use as much 5′ flank sequence as possible from the cloned genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. Accordingly, 1.4 kb, 2.4 kb, 2.4 kb and 3.0 kb of 5′ flank sequence were used to generate *lacZ* fusion constructs for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively. However, it should be noted that 700 bp of the 3.0 kb fragment used for the *cpr-6::lacZ* construct are derived from the original pJB8 cosmid vector. Since this 700bp region is located at the most distal end of the 3.0 kb fragment (Chapter 3, Section 3.2.12), there is a maximum of 2.3 kb of 5′ flank sequence from the *cpr-6* gene available for use with *lacZ* fusion constructs. Of this 2.3 kb, at least 1.35 kb of 5′ flank sequence most proximal to the *cpr-6* gene are thought to be contiguous with the *C.elegans* genome from diagnostic digests of *C.elegans* genomic DNA (Chapter 3, Section 3.2.12).

The second consideration was the amount of coding sequence to include in each construct. The cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* all possess putative signal peptides. Therefore, it was possible that the use of too much coding sequence might result in the amino terminal prepeptide overriding the nuclear localisation signal of the *lacZ* fusion vector. Thus, I decided to use the minimum amount of coding sequence, resulting in no more than eight amino acid residues of the endogenous gene being coded for in the *lacZ* fusion. Three genes, *cpr-3*, *cpr-4* and *cpr-6*, possess introns either immediately after or very close to the ATG translation initiation codon. For these three genes, the first intron was also included in the fusion transgene. These introns were included because potential regulatory elements have previously been observed in the first intron of *cpr-1* (Ray and McKerrow, 1992). Though no such regulatory elements were identified in the first introns of *cpr-3*, *cpr-4* and *cpr-6*, these introns were included as an added precaution.

5.3.6. Vector construction

A summary of the general approach used to obtain the fragments necessary to generate in-frame translational fusions of the 5′ flank regions of each gene with the *lacZ* gene of the pPD21.28 or pPD95.03 vectors is shown in Figure 5.18. For each gene, two PCR primers, X and Y were designed. Primer 'X' was designed to anneal upstream of the unique restriction site 'U' for each gene. These restriction sites (marked 'U') were identified in the sequenced region of the 5′ flank of each gene (Figure 5.19). The sites were subsequently shown to be unique within the cloned 5′ flanks of each gene by digestion of the subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+, containing the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* respectively (see Chapter 3, Figures 3.25, 3.26, 3.27 and 3.28), with the appropriate restriction endonuclease. Primer 'Y' was designed to link the coding region of each gene in frame with the *lacZ* gene of the *lacZ* fusion vectors. Therefore, for each gene, primer 'Y' was designed to anneal to the coding region downstream of the ATG translation initiation codon, and an appropriate restriction endonuclease site 'A' was engineered into the 5′ flank of the primer to facilitate cloning into the polylinker of the *lacZ* fusion vectors. The PCR primers 'X' and 'Y' were used to amplify the intervening sequences for each gene, using the heat stable proof-reading *Pwo* DNA polymerase and a total of only fifteen cycles of PCR to minimise the number of mutations introduced into the amplified region. The sequences of the 'X' and 'Y' primers for each gene are given in (Chapter 2, Section 2.15). The sequences of the 'Y' primers are also shown in Figure 5.20, along with the region of each gene to which these primers anneal. The amplification products were then digested with the appropriate restriction endonucleases recognising sites 'U' and 'A' to generate 75-300bp fragments for linking the bulk of the 5′ flank of each gene to the *lacZ* gene.

The bulk of the 5′ flank region of each gene was obtained by restriction endonuclease digestion of the subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+, with the appropriate restriction endonucleases to release the cloned 5′ flank sequences. In all four cases, the subclones were digested with the restriction endonuclease that recognised the unique restriction site 'U' in the sequenced region of the 5′ flank of each of the four genes (Figure 5.18). The subclones were digested at site 'B' (Figure 5.18), either with the restriction endonuclease originally used to subclone the inserts (CL#034/1.7/KS+, ZK1055/2.8/KS+ and C25B8/3.8/KS+) or with a restriction endonuclease present in the MCS of the pBluescript II KS- vector (W02B2/3/KS+). In the latter case, this was necessary to obtain a fragment that could be cloned into the polylinker of the *lacZ* fusion vector. The resulting fragments were gel purified prior to ligation into the *lacZ* fusion vectors. The restriction endonucleases which recognise sites 'A', 'B' and 'U' that were used in the preparation of these fragments are shown in Table 5.6A. The size of the fragments inserted into the *lacZ* fusion vectors are also shown in Table 5.6B.

5.3.6.1. Screening the PCR amplified linkers for mutations

The PCR amplified regions for *cpr-3*, *cpr-4* and *cpr-6* were digested with the restriction endonucleases that recognised sites 'A' and 'U' and subcloned into the pBluescript II SK- plasmid. The ligation products were transformed into *E.coli* XL1-blue and colonies were screened by colony PCR, using the M13 Forward and Reverse Primers which anneal to sequences flanking the polylinker of the plasmid vector. Plasmid DNA was subsequently prepared from colonies which generated PCR amplification products of the appropriate size. At least three clones for each gene were sequenced using the T3 and / or the T7 PCR primers. The nucleotide sequences of these clones were compared to the 5′ flank sequences of the appropriate genes. This comparison revealed that at least one of the cloned PCR amplified regions for *cpr-3*, *cpr-4* and *cpr-6* shared complete sequence identity to the cloned 5′ flanks of these genes, and therefore that these clones did not possess any PCR generated mutations. These clones were selected for further use. For *cpr-5*, the unique restriction site 'U' was recognised and cleaved by the restriction endonuclease *Bss*HII. The pBluescript II SK- plasmid possesses two *Bss*HII sites flanking the polylinker and therefore it was not possible to clone and sequence the PCR amplified linkers for *cpr-5*.

5.3.6.2. Preparation of the *lacZ* reporter constructs

The approach used to generate all the *lacZ* reporter constructs is summarised in Figure 5.21. In all cases, Fragment 1 of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, previously prepared from subclones CL#034/1.7/KS+, ZK1055/2.8/KS+, W02B2/3/KS+ and C25B8/3.8/KS+ respectively (see Figure 5.18), were co-ligated into the polylinker of the *lacZ* fusion vectors with the appropriate PCR amplified linker region (Fragment 2). For *cpr-3*, *cpr-4* and *cpr-6*, the PCR amplified linker regions (Fragment 2) were released from pBluescript II SK- using the restriction endonucleases that recognise sites 'A' and 'U' (Figure 5.18) and were gel purified. The PCR amplified linker region (Fragment 2) was co-ligated into the polylinker of the *lacZ* fusion vector with the appropriate 5′ flank DNA fragment (Fragment 1). For these three cases, the same restriction endonucleases that recognised sites 'A' and 'B' of the fragments to be cloned were also used to digest the polylinker of the *lacZ* fusion vector. For *cpr-5*, a different approach was required since the PCR amplified linkers could not be cloned into pBluescript II SK- as an intermediate step and therefore had not been sequenced to check for mutations. Thus the PCR amplified products (Fragment 2) for this cloned gene were digested with the restriction endonucleases that recognise sites 'A' and 'U' (Figure 5.18) and were cloned directly into the polylinker of the pPD21.28 *lacZ* fusion vector, with the cloned 5′ flank of this gene (Fragment 1). The polylinker of pPD21.28 did not possess the *Xho*I site at 'B' required for ligation to the cloned 5′ flank of *cpr-5* (Fragment 1) and was therefore digested with *Sal*I which produces ends compatible with those generated by *Xho*I.

5.3.6.3. Checking the integrity of the *lacZ* reporter constructs

E.coli XL1-blue, transformed with the products of the ligations described above. were plated onto L-agar supplemented with 100μg/ml Ampicillin and the resulting transformants screened using colony PCR. The colony PCR used two primers which annealed to the *lacZ* fusion vector sequences flanking the polylinker, the M13 Reverse primer and the *lacZ*/2 primer or the SYNINT primer (the relative position and orientation of each of these primers is shown schematically in Figure 5.21). The nucleotide sequences of these primers are given in Chapter 2, Section 2.11.3.3. This screen helped determine which clones contained the correct fragments and the correct number of fragments ligated into the *lacZ* fusion vectors. However, the sizes of the PCR products produced were sufficiently large in some cases, to make it hard to determine how many copies of the small PCR linker had also been subcloned into the vectors. Therefore, colonies that generated PCR products of the correct size were further analysed. Clones containing the *lacZ* fusion constructs for *cpr-3*, *cpr-4* and *cpr-6* were analysed by repeating the colony PCR but using the PCR primer 'X' for each gene, instead of the M13 Reverse primer. This would reveal whether the 5′ flank DNA fragment was ligated into the vector in the correct orientation with a PCR linker of the correct size, since a PCR product would only be generated if the 5′ flank fragment was in the correct orientation and the size of the product would be dependent on the size of the PCR linker. For *cpr-6* the NUCLOC primer was also used instead of the *lacZ*/2 primer because it was closer to the polylinker. The position and orientation of the NUCLOC primer is shown schematically in Figure 5.21. The PCR linker for *cpr-6* is only 75 bp and therefore I decided that using a primer more proximal to the polylinker than *lacZ*/2 would help to more accurately determine the size of the PCR linker. Together, the colony PCR results indicated that several clones for each of these constructs possessed a single copy of the 5′ flank DNA fragment in the correct orientation and a single copy of the correct PCR linker cloned into the *lacZ* fusion vector. This approach was not used to screen the clones isolated for the *lacZ* fusion with *cpr-5* because regions of these clones would require sequencing (since the PCR linkers had not been sequenced at an intermediate stage). Thus plasmid DNA was prepared from three colonies which generated PCR products of the correct size in the original colony PCR screen. The DNA was sequenced using primer 'X' for *cpr-5*, the SYNINT primer and the M13 Reverse primer. The sequencing, together with the initial colony PCR data, revealed that one of the selected clones contained a single copy of the correct 5′ flank fragment and a single copy of the correct PCR linker, with no PCR generated mutations, all ligated in the correct orientation into the pPD21.28 *lacZ* fusion vector. Furthermore, the sequencing revealed that no mutations had been introduced during the ligation step. Since the selected clones for the *cpr-3*, *cpr-4* and *cpr-6* *lacZ* fusion constructs were not sequenced during this stage, plasmid DNA was prepared from overnight cultures of these and digested with the restriction endonucleases that recognised sites 'A', 'B' and 'U' to ensure that the ligations had not resulted in mutations at these sites. The results of this indicated that all the restriction endonuclease sites utilised in the cloning strategies could be recut and therefore were not mutated.

5.3.6.4. Final analysis of the *lacZ* reporter constructs prior to injection

Medium or large scale plasmid DNA preparations were made from the chosen clones containing the constructs for each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. A sample of each construct was subjected to a single round of sequencing using either the NUCLOC or SYNINT primers. This sequencing confirmed that the restriction enzyme site 'A' for each gene was intact and that translational gene fusions had been generated (Figure 5.22). The results of the previous colony PCR screens, diagnostic digests and sequencing reactions indicated that the generation of the reporter constructs had been successful, with the correct number and size of fragments being cloned into each of the *lacZ* fusion vectors, and no evidence of the introduction of mutations during these steps. The chosen constructs containing the 5′ flanks of *cpr-3*, *cpr-4* and *cpr-6* cloned into pPD95.03 were named *cpr-3*::*lacZ*, *cpr-4*::*lacZ* and *cpr-6*::*lacZ* respectively. The chosen construct containing the 5′ flank of *cpr-5* cloned into pPD21.28 was named *cpr-5*::*lacZ*. The constructs *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ*, containing 1.4kb, 2.4kb, 2.4kb and 3.0kb of 5′ flank sequence of the appropriate gene respectively, were subsequently used to generate transgenic lines of *C.elegans*.

5.3.7. Generating *C.elegans* lines transgenic for the reporter constructs and staining for ß-Galactosidase activity

The syncytial Gonads of N2 wild type young adult hermaphrodite worms were microinjected with 200ng/µl of the desired construct and 100ng/µl of the pRF4 marker plasmid, containing the dominant *rol-6(su1006)* collagen mutation (Mello *et al.,* 1991). This mutation results in a helically twisted cuticle which causes the worm to roll over about its longitudinal axis and to move in circles (the roller phenotype). F1 progeny displaying the roller phenotype were selected and 10 - 20 of these animals seeded onto each of several fresh plates. Single F2 progeny displaying the roller phenotype were subsequently selected from these plates and seeded onto fresh plates. F2 progeny yielding 10% - 90% offspring with the roller phenotype were selected as transmitting lines. The microinjections produced four, five, six and two transmitting lines for the constructs *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ*, respectively.

For staining, 10-20 larval and adult worms with the roller phenotype were seeded onto fresh plates and cultured until the bacterial lawn was almost cleared. The worms were then fixed with methanol and acetone and stained for ß-galactosidase activity using X-gal as a substrate (Chapter 2, Sections 2.17.4 and 2.17.5).

5.3.8. Gut-specific expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The tissue-restricted patterns of expression generated by the four reporter constructs *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* are all remarkably similar (Figures 5.23, 5.24, 5.25 and 5.26). Furthermore, for all four constructs, this tissue-restricted pattern of expression was reproducible between different lines generated from the same construct. In all cases, large stained nuclei were localised along either side of the intestine, posteriorly of the pharyngeal-intestinal valve and anteriorly of the anus.

The entire intestine is derived exclusively from a single founder cell, the E cell, which gives rise to no other tissue. The daughter cells of E are the first to enter the body cavity during gastrulation, 90 minutes after the first cleavage. By 300 minutes, these cells have formed two rows, either side of the developing gut. The anterior most cells of each row subsequently divide dorso-ventrally to give rise to the four cells of int1, which attaches anteriorly to the pharyngeal-intestinal valve. The remaining intestinal cells undergo a 90° left hand twist (Sulston *et al.*, 1983). This twist subsequently increases to 180° and is thought to be a result of packing of the posterior nuclei. A schematic diagram of part of the intestinal tract, demonstrating these features is shown in Figure 5.27. Thus, by hatch, the intestine is comprised of a hollow tube, made up of twenty mononucleate cells which can be divided into nine units along its length (int1 to int9). The lumen of the intestine is surrounded by the four cells of int1 at the anterior end, and subsequently by pairs of cells that make up int2 to int9. Fifteen minutes after the beginning of the L_1 lethargis, the nuclei of the intestinal cells divide. Normally the six anterior most nuclei (of int1 and int2) do not divide, and any of the four posterior most nuclei (of int8 and int9) may also fail to divide. Thus between 10 and 14 of the intestinal nuclei divide, without cytokinesis, to produce between 30 and 34 gut nuclei in twenty cells (Sulston and Horvitz, 1977). All the nuclei of the intestine, including those that do not demonstrate nuclear division, undergo an endoreduplication event during lethargis of each larval moult. This event occurs without chromosome condensation and results in all the cells of the gut having a ploidy of 32C by the final moult (Hedgecock and White, 1985).

The features of the gut nuclei described above make the task of distinguishing the intestine from other tissues relatively simple. Thus the position, distribution and large size of the staining nuclei obtained with all the transgenic lines generated by the four reporter constructs, *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* indicate that they belong to intestinal cells (Figures 5.23, 5.24, 5.25 and 5.26, respectively). Furthermore, though individual intestinal cells could not be discerned, pairs of stained nuclei were frequently observed. These probably represent the binucleated cells of the intestine. These results suggest that expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* is restricted specifically to the intestinal cells of *C.elegans*. Indeed, the expression patterns generated by *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* are very similar to those observed with *lacZ* fusions of two P-glycoprotein genes, *pgp-1* and *pgp-3*, both of which show tissue-restricted expression in the intestinal cells of *C.elegans* (Lincke *et al.*, 1993). At this point, it should be noted that the nuclear localisation of the ß-galactosidase gene is a result of the nuclear localisation signal of the *lacZ* fusion vectors used, and therefore is not representative of the subcellular localisation of the products of the endogenous copies of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*.

5..3.8.1. Variable expression of the *lacZ* fusion transgenes in the intestine

For all the lines generated from the four *lacZ* constructs for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, tissue-restricted staining of only the gut cell nuclei was consistent. However, there was a significant degree of variation in the number of stained nuclei, and the staining intensity. This variation was not only observed between lines generated from different constructs and between lines generated from the same constructs but also between different animals from the same line. The type of variation observed for all four transgene fusions is consistent with mosaicism resulting from the transgenes being maintained as extrachromosomal arrays (discussed in Section 5.3.1). However, other factors could also account for the observed variation. First, the variation could be a result of fixation artefacts, with incomplete fixation resulting in less than the full complement of intestinal cells being equally available for staining. Second, the variation could be a result of the limits of sensitivity of the staining method used. Thus, some intestinal cells may not be expressing the fusion transgene at sufficient levels for detection.

5.3.8.2. Variability in the number and intensity of staining nuclei may be due to the different *lacZ* fusion vectors used

One of the most notable differences in the staining patterns observed with the different constructs was in sensitivity. Thus the gut cell nuclei of all the transgenic lines generated from constructs utilising the pPD95.03 *lacZ* fusion vector (*cpr-3*::*lacZ*, *cpr-4*::*lacZ* and *cpr-6*::*lacZ*) produced staining after 10min incubation with the X-gal substrate. In contrast, lines generated from constructs generated from the pPD21.28 *lacZ* fusion vector (*cpr-5*::*lacZ*) frequently took up to 72 hours to produce visibly stained nuclei using the same fixation and staining procedures, and the frequency of staining worms was also much lower. There was also a marked difference in the degree of mosaicism observed between lines generated using the pPD95.03 *lacZ* fusion constructs and those generated using the pPD21.28 vector. Up to 32 staining gut-cell nuclei were observed relatively frequently in larval and adult stages of lines containing the constructs *cpr-3*::*lacZ*, *cpr-4*::*lacZ* and *cpr-6*::*lacZ*. In contrast, no more than 20 staining gut-cell nuclei were observed for lines containing the construct *cpr-5*::*lacZ*. These observed differences may be a result of the different *lacZ* fusion vectors used to generate the different constructs. Thus all lines generated from different constructs utilising the same pPD95.03 vector produced more rapid and more frequent staining, in addition to lower levels of mosaicism, than lines derived from constructs utilising the pPD21.28 vector. The pPD95.03 vector is a member of the second generation of *lacZ* fusion vectors, produced by A.Fire, which contain introns in the *lacZ* coding region.

These vectors exhibit sensitivity two to three orders of magnitude greater than the older vectors such as pPD21.28 when assayed using the *unc-54* promotor and also exhibit reduced levels of mosaicism with certain promotor sequences (A.Fire, pers.comm.),. Therefore, the features of the new generation of *lacZ* fusion vectors could explain the differences observed. In turn, this suggests that the variation observed is not a result of 'true' mosaicism (i.e. loss of the transgene array in some intestinal cells) but rather a result of insufficient levels of expression of the transgene for detection in certain cells. Towards confirming this observation, it would be necessary to generate constructs with the same 5′ flank sequences fused to different *lacZ* fusion vectors.

5.3.8.3. *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be expressed in different regions of the intestine

Despite the variation observed with the staining patterns for all four gene fusions, resulting from mosaicism or other factors, there was some evidence of differential expression of the four gene fusions within the intestine. For example, a comparison of the expression patterns obtained with the *cpr-3*::*lacZ* and *cpr-4*::*lacZ* transgene fusions reveals some interesting differences. For *cpr-3*::*lacZ*, the majority of intestinal cell nuclei exhibited strong staining for ß-galactosidase activity at two distinct regions in adult worms, the anterior-most end of the intestine and the region spanning the mid-posterior end of the intestine (Figure 5.23A). For *cpr-4*::*lacZ*, the majority of adults exhibited equally strong staining of intestinal cell nuclei along the entire length of the gut (Figure 5.24A). For *cpr-3*::*lacZ*, there was also some evidence of a switch in the pattern of expression occurring during larval development. The majority of L_1 and L_2 larvae exhibited intense staining of intestinal cell nuclei spanning virtually the entire length of the intestine while the majority of L_3 and L_4 larvae exhibited the more polarised staining pattern observed with adult stages (Figure 5.23). In contrast, for *cpr-4*::*lacZ*, no staining pattern was more predominant than any other for the larval stages.

These observations suggest that it is possible that the cathepsin B-like genes of *C.elegans* may be differentially expressed along the length of the intestine and, in some cases, may be expressed in different cells of the gut during *C.elegans* development. It is worth noting that P.MacMorris and T.Blumenthal (pers.comm.) have observed evidence of a conserved pattern in the appearance of gut cell nuclei expressing the *vit-2* gene in adult worms, using *lacZ* gene fusions. Initially, nuclei of the mid and posterior gut were seen to express *vit-2*, followed by additional expression at a third focus at the anterior of the gut. Later, expression extended from each focus to include the surrounding cells until all nuclei of int2 to int8 stained. However, as a result of the mosaic-like variability observed for the gene fusions reported here, and the fact that no one expression pattern was exclusively associated with a given *lacZ* fusion construct, any conclusions regarding expression patterns restricted to certain regions of the gut, or patterns which vary with the development of the worm, can only be made with caution.

5.3.9. The temporal expression patterns indicated from the *lacZ* fusion analysis correlate with those indicated by s-q rtPCR

As discussed in Section 5.3.1, all the *lacZ* fusion constructs are maintained as extrachromosomal arrays in the transgenic worms. Since the arrays are not linked to chromosomes, they may not segregate properly during mitosis or meiosis. This results in transgenic worms not only exhibiting mosaicism in somatic tissues but also transmitting the transgenes in a non-Mendelian manner. Thus, it was not possible to accurately score the proportion of worms that stained for each developmental stage, since there was no way of determining whether lack of staining was due to lack of expression from an array or due to absence of an array. Accordingly, for each gene, accurate comparisons between the relative transcript abundance (determined by s-q rtPCR analysis) and the relative frequency with which worms expressed the appropriate transgene at different stages of development could not be made. Nevertheless, surveys of stained populations of worms did reveal several strong trends that did not require such accurate analysis and these trends are in agreement with the temporal expression patterns of the four genes determined using s-q rtPCR analysis.

In essence, the s-q rtPCR data suggest that induction of expression of *cpr-3*, *cpr-4* and *cpr-5* occurs around the time of embryo hatch. The results obtained with the *lacZ* fusions suggest that induction of expression of *cpr-3*, *cpr-4* and *cpr-5* may occur during late embryogenesis rather than at hatch. This apparent discrepancy is probably due to the use of a mixed stage embryo RNA sample with the s-q rtPCR (and Northern blot) analyses. The transcript abundance of *cpr-3*, *cpr-4* and *cpr-5* would be under-represented in such a sample if these genes are only expressed in the later stages of embryogenesis, as suggested by the *lacZ* data. Such an under-representation would give rise to an artificially perceived increase in relative transcript abundance between the embryo and L_1 stages for each of these genes using s-q rtPCR and would result in the conclusion that induction of expression of these three genes was occurring at the time of hatch rather than during late embryogenesis.

The s-q rtPCR results suggest that *cpr-3*, *cpr-4* and *cpr-5* are all expressed during the four larval and adult stages of *C.elegans* development, albeit with different patterns of relative transcript abundance (Section 5.2.3.5). *C.elegans* strains transgenic for the *lacZ* fusions of *cpr-3*, *cpr-4* and *cpr-5* all stained for ß-galactosidase activity during the four larval and adult stages of *C.elegans* development and therefore support the results obtained from the s-q rtPCR experiments.

The s-q rtPCR data for *cpr-6* produced a temporal pattern of expression very different from the other three genes, with significant levels of expression first occurring around the L_3 stage and increasing through the subsequent L_4 and early adult stages. Lines transgenic for *cpr-6*::*lacZ* showed almost exactly the same pattern of expression as predicted by the s-q rtPCR, the principle difference being that the *lacZ* fusion data suggested that *cpr-6* expression was being induced during the L_2 stage, slightly earlier than predicted from the graphical representation of the s-q rtPCR data. Thus the intestinal cell nuclei of L_3, L_4 and adult stages stained at a high frequency while those of L_2 stage larvae stained only at a low frequency and neither the L_1 or embryo stages stained at all.

These results demonstrate that both the s-q rtPCR and *lacZ* gene fusion approaches yield similar temporal patterns of expression for each of the four genes. Most notably, both approaches are capable of distinguishing the very different temporal pattern of expression of *cpr-6* from those of the other three genes.

5.4. Discussion

The experiments described in this chapter were performed to determine the temporal and spatial patterns of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. Obtaining information on when and where a gene is expressed can be very helpful in assigning potential biological roles for that gene's product. When such analysis is performed on a family of related genes, the information obtained may also help to determine how the members of the gene family interact at the biological level.

5.4.1. Overview of results obtained

The Northern blot studies indicate that transcription results in a single predominant mRNA species for each of the genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, suggesting that these genes are not alternatively spliced. These results agree with the 5′ RACE experiments (Chapter 4, Section 4.2.6) which indicate that transcription initiation occurs predominantly from one region of the 5′ flank of each of the four genes. The Northern blot studies also suggest that expression of *cpr-4*, *cpr-5* and *cpr-6* is induced at some point at or after embryo hatch (Section 5.2.2), supporting the data obtained for these three genes using s-q rtPCR (Section 5.2.4). Such conclusions cannot be made for *cpr-3* because the probe used to control for equal loading of RNA during Northern analysis failed to generate any signal.

The s-q rtPCR experiments suggest that *cpr-3* is essentially constitutively expressed during the larval and adult stages of *C.elegans* development, but that expression of this gene may be increased slightly during progression from the embryo to L_1 stage. For *cpr-4*, the results suggest that expression of this gene occurs predominantly during the larval stages of development, since transcript abundance is reduced both in the embryo and adult stages. The s-q rtPCR results suggest that *cpr-5* expression is induced at hatch, rises to a maximum at the L_2 stage and shows a slight decline in abundance in subsequent larval and adult stages. For, *cpr-6*, the results suggest that expression of this gene remains low during the embryo and early larval stages of development, but is induced during the L_3 stage and continues to increase through the L_4 and early adult stages. In addition, both *cpr-5* and *cpr-6* exhibit highly developmentally regulated patterns of expression, showing over 15 fold increases in their transcript abundance between minimal and maximal expression. Finally, the temporal expression patterns of the four genes obtained using s-q rtPCR correlate well with the expression data obtained from the appropriate *lacZ* transgene fusion. Most notably, both approaches identified *cpr-6* as having a very different temporal expression pattern from *cpr-3*, *cpr-4* and *cpr-5* and both approaches generated a very similar expression pattern for this gene. Finally, the results obtained using the *lacZ* transgene fusions suggest that all four genes are expressed exclusively in the intestinal cells of *C.elegans*.

5.4.2. Efficacy of the results obtained using *lacZ* reporter gene fusions

The extensive use of *lacZ* gene fusions has demonstrated that this approach is frequently capable of faithfully reproducing the spatial expression patterns of endogenous genes in *C.elegans*. However, this approach, like any other, may produce spurious results in some cases (discussed in Section 5.3.3) and therefore cannot be used as the sole means of determining the physiological expression patterns of endogenous genes. Thus, it will be necessary to confirm the results reported here by another approach, such as *in situ* hybridisation or antibody staining of whole worms. Though, the spatial expression patterns of the *cpr-3*::*lacZ*, *cpr-4*::*lacZ*, *cpr-5*::*lacZ* and *cpr-6*::*lacZ* gene fusions have not as yet been confirmed using an alternative method, there is a substantial amount of indirect evidence to suggest that they do reflect the spatial expression patterns of the endogenous genes.

All four *lacZ* gene fusions produced similar gut-restricted patterns of expression (Section 5.3.8). These results suggest that it is the presence of regulatory elements in the 5′ flanks of these four genes that is responsible for generating the expression patterns observed, since it is unlikely that four different 5′ flank sequences would reproducibly generate the same artefact. Furthermore, these patterns of expression were obtained using two different *lacZ* fusion vectors, pPD95.03 and pPD21.28 which suggests that the gut-restricted expression patterns observed are not a result of the type of vector used.

Ectopic expression generally results in a few additional cells staining in tissues outwith those which express the endogenous gene. Therefore, in cases where ectopic expression is occurring, one expects to observe staining in more than one tissue. The ability of all four *lacZ* gene fusions reported here to produce staining only within the intestinal cells therefore argues against ectopic expression occurring.

The temporal expression patterns of the four genes obtained using both s-q rtPCR and *lacZ* fusion transgene approaches correlate well (Section 5.3.9). The ability of two very different approaches to generate similar temporal expression patterns for each of the four genes provides good evidence in itself for the efficacy of both these approaches. The ability of all four *lacZ* fusion transgenes to reflect the temporal expression patterns of their endogenous genes suggests that all the regulatory elements required to drive appropriate temporal expression of these genes are present in the *lacZ* fusion constructs. In turn, these results suggest that the elements required to drive appropriate spatial expression of the four genes are also present within each construct. Accordingly, one would not expect to be able to reproduce the correct temporal expression pattern of a gene if it were being expressed in inappropriate tissues.

It is commonly recognised that the 3′ UTRs of some mature mRNA transcripts may be important for regulating the stability of those transcripts (Chapter 4, Section 4.3.4.1). This may affect the expression pattern of a gene by stabilising the transcript in certain tissues or at certain times during development. None of the four gene fusions were made using their 3′ flank sequences, but instead the *unc-54* 3′ end was used. Therefore it is possible that the temporal and/or spatial expression patterns of the *lacZ* fusion transgenes might be affected by the lack of appropriate 3′ flank sequence. However, again the correlation between the temporal expression patterns obtained using the s-q rtPCR and *lacZ* fusion approaches argue against this. The s-q rtPCR approach provided information on the relevant transcript abundance of each of the four genes from all tissues. These results correlated with the frequency with which the intestinal cells from different stages of *C.elegans* development exhibit staining, in lines transgenic for the appropriate construct. Thus, one would expect any alteration in the temporal and/or spatial expression patterns caused by the use of a different 3′ end in the *lacZ* gene fusions to reduce the correlation observed with the s-q rtPCR approach.

One of the strongest pieces of evidence for the efficacy of the *lacZ* fusion results comes from data obtained previously for *cpr-1* (Ray and McKerrow, 1992). The authors demonstrated that expression of this cathepsin B-like gene occurs exclusively in the gut using *in situ* hybridisation. More recently, this expression pattern has been confirmed using a *lacZ* gene fusion containing only the 5′ flank sequences of *cpr-1* (C.Britton, pers. comm.). Such data suggests that cathepsin B-like genes are localised to the intestinal cells of *C.elegans*. Interestingly, the data also suggest that the absence of a 3′ flank sequence in the *cpr-1::lacZ* gene fusion does not result in alteration of the tissues where the gene is expressed, suggesting that the 3′ UTR of *cpr-1* may not be necessary for appropriate spatial expression of this gene. There is also circumstantial evidence for expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* in the gut of *C.elegans*. Analysis of the 5′ flank sequences of *vit-2* (MacMorris *et al.,* 1992) and *ges-1* (Egan *et al.,* 1995) have identified GATA motifs, including the VPE2 sequences, which are essential for the appropriate gut-specific expression of these two genes in *C.elegans*. All four genes analysed here, in addition to *cpr-1*, possess a number of sequences in their 5′ flanks with homology to these GATA motifs, including VPE2 sequences (Chapter 4, Section 4.3.4.2), suggesting that these sequences may be responsible for the gut-specific expression of *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*.

The final evidence for gut-specific expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* comes from the potential roles one might expect these genes to perform. Such roles are consistent with gut-specific expression of these genes and are discussed in the next section. Before discussing these roles, it should be noted that expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* is not synonymous with activity of their enzymes since the DNA sequence data suggest that the enzymes encoded by these genes are translated as precursors which must be processed. Furthermore, endogenous enzyme inhibitors may further restrict the spatial and temporal distribution of active enzyme translated from each of these genes. For instance, though the *lacZ* transgene fusion data suggest that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are all expressed in the intestinal cells, the intestinal cells may possess different inhibitors resulting in different intestinal cells possessing distinct active cathepsin B-like enzymes.

The analysis of the temporal and spatial expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* suggests that these genes are highly regulated at the transcriptional level. This suggests that the distribution of the enzymes encoded by these genes will be primarily regulated at the transcriptional level. Accordingly, endogenous inhibitors and enzyme processing events are unlikely to result in significant differences between the expression pattern of the gene and the distribution of its active cathepsin B-like enzyme, but may fine tune the distribution of active enzyme once expressed. Such fine tuning may well be at the cellular level, possibly with some active enzymes being targeted to the lysosomes and others being secreted. Indeed, the cathepsin B-like enzyme encoded by *cpr-5* does not possess an N-glycosylation signal suggesting that it may not be targeted to the lysosomes but secreted instead (Chapter 4, Section 4.3.3.2).

5.4.3. Potential biological roles for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The data obtained from the *lacZ* fusions of *cpr-3*, *cpr-4* and *cpr-5* suggest that these genes are expressed exclusively in the intestinal cells of *C.elegans* (Section 5.3.8). These data, in conjunction with the s-q rtPCR data (Section 5.3.9), also suggest that expression of these genes is induced prior to hatch during the threefold stage of embryogenesis and that expression continues in all subsequent larval and adult stages of development, albeit at different levels. Such a temporal and spatial expression profile is consistent with these genes having a role in digestion in the gut of *C.elegans*. Vertebrate cathepsin B is thought to play an important role in protein turnover and degradation within the lysosomes (Chapter 1, Section 1.7.1) but may also play an important role as a precursor processing enzyme (Chapter 1, Section 1.7.2). Since immunohistochemical studies have detected cathepsin B in many human tissues including the epithelia of the gastro-intestinal tract (Howie *et al.,* 1985), it is possible that this enzyme performs an important, but indirect, function in digestion (possibly by processing proteins involved in digestion or by regulating the turnover of proteins involved in digestion). Vertebrate cathepsin B is not secreted into the vertebrate gut lumen and thus is unlikely to perform a direct role in digestion. Therefore, it is also possible that the cathepsin B-like enzymes of *C.elegans* may play an analogous indirect role in digestion within the intestinal cells.

Alternatively, these enzymes may play a more direct role in digestion. Several parasitic nematode and trematode species have been shown to release cysteine protease enzymes into the media when cultured *in vitro* (Chapter 1, Section 1.8) and these enzymes are thought to represent excreted or secreted products. Such studies have suggested that two parasitic trematode species, *F.hepatica* and *H.cylindracea*, excrete/secrete cathepsin B-like enzymes into the media when cultured *in vitro*. Indeed, active cathepsin B-like activity has been identified in the secretory granules of intestinal cells of *F.hepatica* (Yamasaki *et al.,* 1992). More evidence of a direct role for cathepsin B-like enzymes in digestion comes from the trematode *S.mansoni*. *S.mansoni* uses haemoglobin digestion as the principle mechanism for obtaining essential amino acids and the major haemoglobinase of this trematode species (Sm31) has a cathepsin B-like activity (Gotz and Klinkert, 1993). Together, these data suggest that cathepsin B-like enzymes may be secreted into the intestinal lumen and therefore may be directly involved in digestion in nematode and trematode species. To date, cathepsin B-like multigene families have been identified in the parasitic nematode species, *H.contortus* and *O.ostertagi* (Chapter 1, section 1.9) in addition to the multigene family of *C.elegans* reported in this thesis. There is also some evidence for the presence of such a multigene family in the parasitic nematode species, *A.caninum* and the parasitic trematode species, *F.hepatica* (Chapter 1, Sections 1.9). These data suggest that cathepsin B-like multigene families may be common to many nematode and trematode species. This is in agreement with a direct role in digestion, since amplification of this gene family may have been required to produce several enzymes which, when acting together, can degrade the broad spectrum of substrates necessary for a direct role in digestion.

The *cpr-6* gene is also expressed in the intestinal cells of *C.elegans* and therefore may also perform a digestive role in conjunction with *cpr-3*, *cpr-4* and *cpr-5*. However, the temporal expression pattern of this gene suggests that induction of expression does not occur until the later larval stages of development. Such an expression pattern is not expected for a gene involved in digestion since early larval stages must also feed. This does not preclude *cpr-6* having a role in digestion but merely indicates that this gene is not required for digestion in early larval stages. Though *cpr-6* may perform a role in digestion, the unusual temporal expression pattern of this gene compared to *cpr-3*, *cpr-4* and *cpr-5* suggests that it may perform other related roles.

The primary function of the intestine is probably to secrete digestive enzymes and absorb the processed nutrients. However, the intestine of *C.elegans* also represents one of the main storage organs of the body, possessing numerous different storage granules in the cytoplasm of intestinal cells. These storage granules are thought to contain lipids, proteins and carbohydrates (White, 1988). Thus, the cathepsin B-like enzyme encoded by the *cpr-6* gene may be involved in turnover and degradation of stored proteins. Alternatively, this enzyme may be involved in processing stored precursor proteins required during later stages of development. The yolk proteins represent one such family of proteins which may require processing during later stages of *C.elegans* development.

These proteins are synthesised in the intestine of *C.elegans* (Kimble and Sharrock, 1983). The genes encoding these proteins are expressed exclusively in the intestinal cells of hermaphrodite worms, with expression starting at the late L_4 stage and continuing throughout the lifespan of the adult hermaphrodite (Blumenthal *et al.,* 1984). Since the yolk proteins of other organisms are synthesised as precursor proteins that require subsequent processing (Tata, 1976), the yolk proteins of *C.elegans* may also require processing in the intestine of adult hermaphrodite worms. *cpr-6* shows the appropriate temporal and spatial expression pattern for a potential role in yolk protein processing, being expressed in the intestine in the later larval stages of development prior to yolk protein gene expression. It is therefore tempting to speculate that this gene may be required for yolk protein processing in hermaphrodite worms. Since yolk proteins are expressed in a sex specific manner, the enzymes required to process these proteins may also be expressed in a sex specific manner. Therefore, as a first step towards investigating the possibility of *cpr-6* being involved in yolk protein processing, it would be useful to determine whether or not *cpr-6* is expressed in male worms. Since, if *cpr-6* was also found to be expressed exclusively in the intestinal cells of hermaphrodite worms, such findings would provide good circumstantial evidence of a role for this enzyme in yolk protein processing.

.

5.4.4. Biological interaction

The data reported in this chapter suggest that all four genes are expressed in the same tissue and possess overlapping patterns of temporal expression. These data suggest that the members of this gene family may encode proteins which act together to effect their biological roles. However, though the temporal expression patterns overlap, they are clearly different. This may result in different combinations of active cathepsin B-like enzymes at different stages of development. The possibility that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* encode cathepsin B-like enzymes with different activities and substrate specifities (Chapter 4, Section 4.3.3.1) may be important in this respect since different combinations of enzymes may degrade or process very different combinations of protein substrates. Thus biological interaction between the products of this gene family could be very complex, with different enzymes interacting at different stages, possibly to effect different but related functions. For example, one enzyme may be capable of performing both digestive and processing roles but the type of role performed may be dependent on interaction with different members of the gene family. Accordingly, the possible digestive roles for *cpr-3*, *cpr-4* and *cpr-5* and the possible processing role for *cpr-6* are not mutually exclusive, since some or all of these enzymes may interact at various stages of development to effect different but related biological functions.

Since the data suggest that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be expressed simultaneously in the same tissue at certain stages of *C.elegans* development, it is possible that some members of this gene family may be functionally redundant. It is unlikely that any one of the four genes reported here, or *cpr-1*, will be completely functionally redundant (i.e. encode an enzyme which shares identical activity and substrate specificity with another enzyme), since these five genes encode enzymes with diverged predicted amino acid sequences which may have different activities and substrate specificities. It is possible that additional, as yet unidentified, cathepsin B-like genes do exist in the *C.elegans* genome which have identical activities and substrate specificities to the enzymes encoded by the five genes isolated to date. However, this is unlikely since the Southern blot data presented in Chapter 3, Section 3.2.2 does not provide any evidence for the presence of additional cathepsin B-like genes with significant homology to *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* within the genome of *C.elegans*. Thus, if there is any functional redundancy, it is most likely to be a result of the ability of a combination of cathepsin B-like enzymes to mimic the activity and substrate specificity of another individual enzyme. Accordingly, though individual members of the cathepsin B-like gene family may play important roles in *C.elegans* biology, null mutations in individual genes may not result in lethal phenotypes because other members of the gene family, acting together, may be able to assume the biological function of the mutated gene. The presence of functionally redundant members within the cathepsin B-like multigene family would suggest that this gene family is essential for *C.elegans* biology. In order to test such a hypothesis, it would be necessary to obtain *C.elegans* strains each possessing a null mutation of a different member of the cathepsin B-like gene family. These strains could then be crossed to produce multiply mutant strains. Analysis of the phenotypes of strains with mutations in different combinations of cathepsin B-like genes might then reveal which genes play biologically interchangeable roles.

Table 5.1

The table shows the radioactivity data obtained from s-q rtPCR analysis of *cpr-3* using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the *cpr-3* and *ama-1* transcripts were excised using the autoradiographs shown in Figure 5.11 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of *cpr-3* at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the *cpr-3* transcripts to the average cpm obtained from the PCR products derived from the *ama-1* transcripts. The mean and standard deviation of the *cpr-3* relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.

		cpm from PCR product of *cpr-3* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}3}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 1	Embryo	538	559	548.5	1612	1567	1589.5	0.345077
	L1	550	539	544.5	505	491	498	1.093373
	L2	1078	1107	1092.5	893	952	922.5	1.184282
	L3	952	916	934	1220	1254	1237	0.755053
	L4	770	705	737.5	1306	1337	1321.5	0.558078
	Adult	900	925	912.5	1214	1285	1249.5	0.730292
Filter 3	Embryo	194	210	202	192	204	198	1.020202
	L1	141	149	145	94	93	93.5	1.550802
	L2	201	190	195.5	124	96	110	1.777273
	L3	234	210	222	166	137	151.5	1.465347
	L4	151	146	148.5	172	176	174	0.853448
	Adult	171	186	178.5	130	137	133.5	1.337079
Filter 5	Embryo	616	621	618.5	1736	1689	1712.5	0.361168
	L1	401	423	412	344	277	310.5	1.326892
	L2	385	339	362	446	430	438	0.826484
	L3	645	649	647	618	672	645	1.003101
	L4	454	430	442	640	625	632.5	0.698814
	Adult	528	584	556	631	641	636	0.874214

mean and standard deviation of $\frac{cpr\text{-}3}{ama\text{-}1}$ ratios from Filters 1, 3 and 5

	mean	std. dev.
Embryo	0.575482	0.385223
L1	1.323689	0.228731
L2	1.26268	0.480218
L3	1.0745	0.36049
L4	0.703447	0.14774
Adult	0.980528	0.317056

		cpm from PCR product of *cpr-3* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}3}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 2	Embryo	244	251	247.5	356	369	362.5	0.682759
	L1	355	309	332	204	229	216.5	1.533487
	L2	351	361	356	289	283	286	1.244755
	L3	623	578	600.5	488	517	502.5	1.195025
	L4	518	527	522.5	733	752	742.5	0.703704
	Adult	430	402	416	551	515	533	0.780488
Filter 4	Embryo	217	242	229.5	342	320	331	0.693353
	L1	174	192	183	108	129	118.5	1.544304
	L2	582	528	555	275	246	260.5	2.130518
	L3	524	568	546	296	315	305.5	1.787234
	L4	416	455	435.5	378	349	363.5	1.198074
	Adult	468	440	454	482	417	449.5	1.010011
Filter 6	Embryo	627	585	606	1483	1444	1463.5	0.414076
	L1	602	673	637.5	412	406	409	1.55868
	L2	606	598	602	548	565	556.5	1.081761
	L3	1124	1163	1143.5	788	755	771.5	1.482178
	L4	890	963	926.5	888	853	870.5	1.064331
	Adult	1039	1066	1052.5	722	733	727.5	1.446735

mean and standard deviation of $\frac{cpr\text{-}3}{ama\text{-}1}$ ratios from Filters 2, 4 and 6

	mean	std. dev.
Embryo	0.596729	0.158271
L1	1.54549	0.012638
L2	1.485678	0.564363
L3	1.488145	0.29615
L4	0.988703	0.255715
Adult	1.079078	0.338451

Table 5.2

The table shows the radioactivity data obtained from s-q rtPCR analysis of *cpr-4* using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the *cpr-4* and *ama-1* transcripts were excised using the autoradiographs shown in Figure 5.12 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of *cpr-4* at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the *cpr-4* transcripts to the average cpm obtained from the PCR products derived from the *ama-1* transcripts. The mean and standard deviation of the *cpr-4* relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.

		cpm from PCR product of *cpr-4* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}4}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 1	Embryo	177	178	177.5	3690	3710	3700	0.047973
	L1	157	163	160	671	662	666.5	0.24006
	L2	185	164	174.5	1050	982	1016	0.171752
	L3	352	355	353.5	2504	2544	2524	0.140055
	L4	661	583	622	2958	3113	3035.5	0.204909
	Adult	303	326	314.5	3551	3377	3464	0.090791
Filter 3	Embryo	212	245	228.5	4513	4453	4483	0.05097
	L1	391	373	382	1423	1427	1425	0.26807
	L2	215	213	214	1563	1518	1540.5	0.138916
	L3	484	479	481.5	2010	2020	2015	0.238958
	L4	615	580	597.5	2122	2228	2175	0.274713
	Adult	397	402	399.5	3757	3919	3838	0.104091
Filter 5	Embryo	197	244	220.5	4451	4502	4476.5	0.049257
	L1	327	347	337	1677	1591	1634	0.206242
	L2	234	225	229.5	1351	1418	1384.5	0.165764
	L3	440	416	428	2511	2566	2538.5	0.168604
	L4	758	796	777	2603	2610	2606.5	0.298101
	Adult	391	418	404.5	4343	4607	4475	0.090391

		cpm from PCR product of *cpr-4* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}4}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 2	Embryo	236	181	208.5	3526	3563	3544.5	0.058824
	L1	196	266	231	924	921	922.5	0.250407
	L2	158	153	155.5	1006	1121	1063.5	0.146215
	L3	378	381	379.5	2801	2783	2792	0.135924
	L4	632	669	650.5	2932	2978	2955	0.220135
	Adult	334	316	325	3857	3854	3855.5	0.084295
Filter 4	Embryo	216	224	220	3507	3345	3426	0.064215
	L1	273	311	292	890	1021	955.5	0.305599
	L2	203	209	206	1267	1291	1279	0.161063
	L3	410	384	397	1700	1657	1678.5	0.236521
	L4	612	650	631	1962	1902	1932	0.326605
	Adult	370	384	377	3417	3483	3450	0.109275
Filter 6	Embryo	214	178	196	2451	2509	2480	0.079032
	L1	289	303	296	904	905	904.5	0.327253
	L2	227	221	224	1097	1030	1063.5	0.210625
	L3	365	407	386	1733	1796	1764.5	0.218759
	L4	615	618	616.5	1644	1654	1649	0.373863
	Adult	329	323	326	2690	2785	2737.5	0.119087

mean and standard deviation of $\frac{cpr\text{-}4}{ama\text{-}1}$ ratios from Filters 1, 3 and 5

	mean	std. dev.
Embryo	0.0494	0.001504
L1	0.238124	0.030959
L2	0.158811	0.017487
L3	0.182539	0.050903
L4	0.259241	0.048484
Adult	0.095091	0.007797

mean and standard deviation of $\frac{cpr\text{-}4}{ama\text{-}1}$ ratios from Filters 2, 4 and 6

	mean	std. dev.
Embryo	0.067357	0.010464
L1	0.294419	0.039624
L2	0.172635	0.033728
L3	0.197068	0.053692
L4	0.306868	0.078741
Adult	0.104219	0.017938

Table 5.3

The table shows the radioactivity data obtained from s-q rtPCR analysis of *cpr-5* using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the *cpr-5* and *ama-1* transcripts were excised using the autoradiographs shown in Figure 5.13 as templates. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of *cpr-5* at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the *cpr-5* transcripts to the average cpm obtained from the PCR products derived from the *ama-1* transcripts. The mean and standard deviation of the *cpr-5* relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 1, 3 and 5 or from filters 2, 4 and 6.

		cpm from PCR product of *cpr-5* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}5}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 1	Embryo	581	566	573.5	3055	3034	3044.5	0.188372
	L1	1740	1757	1748.5	577	589	583	2.999142
	L2	5246	5237	5241.5	818	872	845	6.202959
	L3	6683	6740	6711.5	1427	1402	1414.5	4.744786
	L4	8590	8926	8758	1511	1471	1491	5.87391
	Adult	6165	6121	6143	1893	1934	1913.5	3.210348
Filter 3	Embryo	417	436	426.5	2780	2847	2813.5	0.151591
	L1	1504	1583	1543.5	550	569	559.5	2.758713
	L2	4815	4758	4786.5	693	659	676	7.080621
	L3	7312	7295	7303.5	1864	1875	1869.5	3.90666
	L4	6583	6584	6583.5	1191	1186	1188.5	5.539335
	Adult	3717	3687	3702	1494	1587	1540.5	2.403116
Filter 5	Embryo	140	151	145.5	304	270	287	0.506969
	L1	1313	1349	1331	336	302	319	4.172414
	L2	4085	4004	4044.5	419	427	423	9.561466
	L3	4432	4600	4516	624	648	636	7.100629
	L4	4338	4443	4390.5	544	542	543	8.085635
	Adult	3045	3126	3085.5	735	720	727.5	4.241237

		cpm from PCR product of *cpr-5* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}5}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 2	Embryo	743	719	731	3191	3089	3140	0.232803
	L1	2437	2501	2469	592	616	604	4.087748
	L2	7109	7216	7162.5	916	875	895.5	7.998325
	L3	9621	9762	9691.5	1832	1838	1835	5.281471
	L4	11156.7	11086	11121	1999	1892	1945.5	5.716523
	Adult	8424	8419	8421.5	2463	2440	2451.5	3.435244
Filter 4	Embryo	532	517	524.5	3398	3432	3415	0.153587
	L1	2290	2206	2248	744	787	765.5	2.936643
	L2	5732	5792	5762	958	977	967.5	5.955556
	L3	7517	7595	7556	1957	1913	1935	3.90491
	L4	8033	8073	8053	1853	1930	1891.5	4.257468
	Adult	5129	5124	5126.5	2103	2169	2136	2.400047
Filter 6	Embryo	178	189	183.5	373	367	370	0.495946
	L1	2808	2957	2882.5	488	499	493.5	5.840932
	L2	7171	6892	7031.5	486	498	492	14.29167
	L3	8273	8310	8291.5	846	816	831	9.977738
	L4	8697	8688	8692.5	907	1020	963.5	9.021796
	Adult	3963	3888	3925.5	1268	1197	1232.5	3.18499

mean and standard deviation of $\frac{cpr\text{-}5}{ama\text{-}1}$ ratios from Filters 1, 3 and 5

	mean	std. dev.
Embryo	0.282311	0.195427
L1	3.31009	0.756408
L2	7.615015	1.74186
L3	5.250692	1.655994
L4	6.499627	1.383673
Adult	3.2849	0.921326

mean and standard deviation of $\frac{cpr\text{-}5}{ama\text{-}1}$ ratios from Filters 2, 4 and 6

	mean	std. dev.
Embryo	0.294112	0.179225
L1	4.288441	1.462509
L2	9.415182	4.344917
L3	6.38804	3.184051
L4	6.331929	2.441055
Adult	3.00676	0.540123

Table 5.4

The table shows the radioactivity data obtained from s-q rtPCR analysis of *cpr-6* using the modified protocol. See Figure 5.10 for a summary of the protocol.

The regions of the six filters corresponding to the PCR amplification products derived from the *cpr-6* and *ama-1* transcripts were excised using the autoradiographs shown in Figure 5.14 as templates. However, the samples taken from filters 3 and 5 were lost and therefore only the data for filters 1, 2, 4 and 6 are shown. The amount of radioactivity (in counts per minute, cpm) from each band was measured twice by counting in scintillant. The average of the two counts was taken. The relative transcript abundance of *cpr-6* at each developmental stage was calculated as a ratio of the average cpm obtained from the PCR products derived from the *cpr-6* transcripts to the average cpm obtained from the PCR products derived from the *ama-1* transcripts. The mean and standard deviation of the *cpr-6* relative transcript abundance ratio at each developmental stage was calculated from the results obtained from three independent PCR reactions. Accordingly, the mean and standard deviation for each developmental stage was calculated from filters 2, 4 and 6.

		cpm from PCR product of *cpr-6* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}6}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 1	Embryo	224	215	219.5	1037	1044	1040.5	0.210956
	L1	470	418	444	374	363	368.5	1.204885
	L2	183	171	177	364	374	369	0.479675
	L3	524	485	504.5	748	861	804.5	0.627098
	L4	2250	2255	2252.5	1110	1134	1122	2.007576
	Adult	5155	5106	5130.5	781	844	812.5	6.314462

		cpm from PCR product of *cpr-6* cDNA			cpm from PCR product of *ama-1* cDNA			Ratio of Av. cpm: $\frac{cpr\text{-}6}{ama\text{-}1}$
		cpm1	cpm2	Av. cpm	cpm1	cpm2	Av. cpm	
Filter 2	Embryo	220	217	218.5	927	988	957.5	0.228198
	L1	622	676	649	508	507	507.5	1.278818
	L2	189	161	175	419	483	451	0.388027
	L3	645	663	654	961	975	968	0.67562
	L4	2321	2330	2325.5	1423	1290	1356.5	1.714338
	Adult	4104	4006	4055	978	971	974.5	4.161108
Filter 4	Embryo	251	219	235	1431	1402	1416.5	0.165902
	L1	180	162	171	272	234	253	0.675889
	L2	581	588	584.5	688	650	669	0.873692
	L3	441	440	440.5	764	747	755.5	0.583058
	L4	2000	1917	1958.5	1060	1085	1072.5	1.826107
	Adult	2451	2384	2417.5	999	994	996.5	2.425991
Filter 6	Embryo	538	511	524.5	4134	4075	4104.5	0.127787
	L1	1067	1061	1064	1934	1933	1933.5	0.550297
	L2	459	519	489	2784	2774	2779	0.175963
	L3	1605	1538	1571.5	4686	4771	4728.5	0.332346
	L4	6174	6079	6126.5	4422	4357	4389.5	1.395717
	Adult	10845.3	10691	10768	3930	3923	3926.5	2.742365

mean and standard deviation of $\frac{cpr\text{-}6}{ama\text{-}1}$ ratios from Filters 2, 4 and 6

	mean	std. dev.
Embryo	0.173962	0.050689
L1	0.835001	0.389452
L2	0.479227	0.357694
L3	0.530341	0.177605
L4	1.645388	0.223326
Adult	3.109821	0.924081

Table 5.5

Statistical analysis of the temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* (shown in Figure 5.16) using Student's *t*-test. For each gene, the relative transcript abundance for each stage was compared to the next stage in development (i.e. embryo and L_1, L_1 and L_2, L_2 and L_3, L_3 and L_4, L_4 and adult). Each *t*-test was performed with three sets of paired data. The three sets of data were derived from the three independent PCR reactions (A, B and C, Figure 10) and each pair comprises the relative transcript abundance ratios of two adjacent developmental stages derived from the same PCR reaction. The probabilities were interpolated from a table of tabulated values of *t* using two degrees of freedom.

The *t*-test null hypothesis assumes that two samples have been derived from the same population, and therefore that any differences in the mean of these samples can be accounted for by sampling errors. Accordingly, the percentage probabilities indicate the probability of the relative transcript abundance ratios of two adjacent developmental stages being derived from the same population. Probabilities of less than or equal to 5% were deemed significant (**S**) and probabilities less than or equal to 1% were deemed highly significant (**HS**).

Table 5.5

	cpr-3	*cpr-4*	*cpr-5*	*cpr-6*
Embryo-L_1	2.70% **S**	0.84% **HS**	1.13% **S**	7.81%
L_1-L_2	81.02%	9.39%	2.08% **S**	37.52%
L_2-L_3	41.72%	60.85%	4.16% **S**	79.78%
L_3-L_4	9.63%	10.94%	2.38% **S**	0.33% **HS**
L_4-Adult	11.54%	2.64% **S**	1.12% **S**	11.21%

Table 5.6A

The table shows the restriction endonucleases that recognise sites ‘A’, ‘B’ and ‘U’ (Figure 5.18) which were used in the generation of the *lacZ* fusion transgene constructs of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*.

Table 5.6B

The table shows the sizes of Fragment 1 and Fragment 2 (see Figure 5.18) after digestion with the restriction endonucleases which recognise sites ‘A’, ‘B’ and ‘U’. The total size of the 5′ flank used for each gene is also indicated.

Note: 700bp of the 3kb 5′ flank fragment used for *cpr-6* is from the original pJB8 cosmid vector. Therefore, a maximum of 2.3kb of *cpr-6* 5′ flank sequence was used in the generation of the *lacZ* fusion transgene construct for this gene.

Table 5.6A

	enzyme used at site 'A'	enzyme used at site 'B'	enzyme used at site 'U'
cpr-3 (subclone CL#034/1.7/KS+)	*Bam*HI	*Hind*III	*Eco*RV
cpr-4 (subclone ZK1055/2.8/KS+)	*Xma*I	*Hind*III	*Kpn*I
cpr-5 (subclone W02B2/3/KS+)	*Xma*I	*Xho*I	*Bss*HII
cpr-6 (subclone C25B8/3.8/KS+)	*Bam*HI	*Pst*I	*Sal*I

Table 5.6B

	Size of Fragment 1 after digeston with enzymes recognising sites 'B' and 'U'	Size of Fragment 2 after digestion with enzymes recognising sites 'A' and 'U'	total size of 5' flank used in the *lacZ* fusion constructs
cpr-3 (subclone CL#034/1.7/KS+)	1.1 kb	300 bp	1.4 kb
cpr-4 (subclone ZK1055/2.8/KS+)	2.2 kb	210 bp	2.4 kb
cpr-5 (subclone W02B2/3/KS+)	2.2 kb	150 bp	2.4 kb
cpr-6 (subclone C25B8/3.8/KS+)	3.0 kb	75 bp	3.0 kb

Figure 5.1

The autoradiographic results of Northern blot analysis obtained after medium stringency washing. Northern blots were prepared from 4µg of total RNA isolated from mixed stage cultures of *C.elegans* (**Lane 1**) and from *C.elegans* embryos (**Lane 2**). The denaturing agarose gel used to generate the Northern blots is shown in Figure 5.3A.

A: Northern blot probed with the cDNA clone of *cpr-3* (cm12b6)

B: Northern blot probed with the cDNA clone of *cpr-4* (cm14e3)

C: Northern blot probed with the cDNA clone of *cpr-5* (cm04d10)

D: Northern blot probed with the cDNA clone of *cpr-6* (cm01a5)

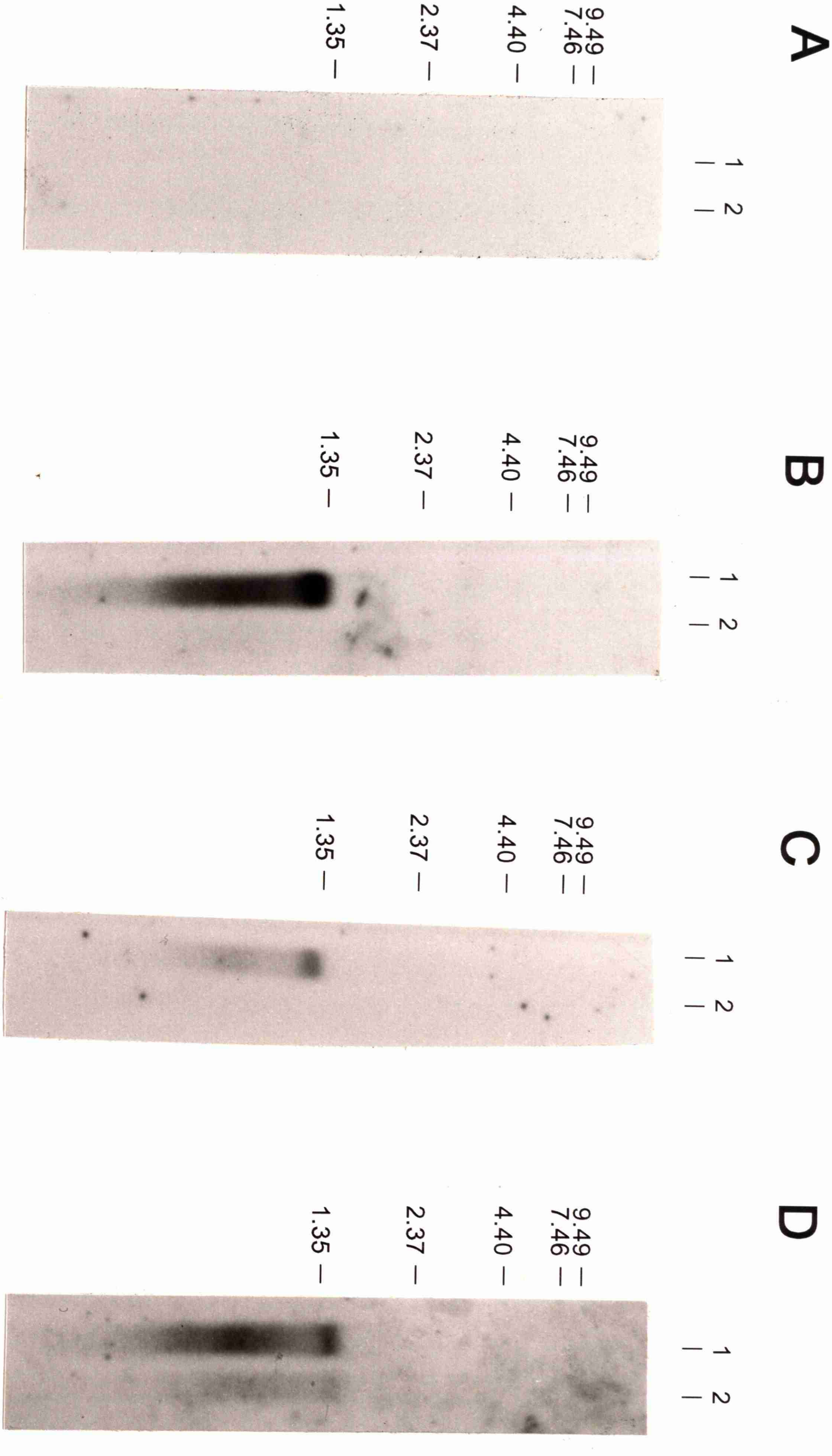

A
B
C
D
9.49
7.46
4.40
2.37
1.35
1
2

Figure 5.2

The autoradiographic results of Northern blot analysis obtained after high stringency washing. Northern blots were prepared from 4µg of total RNA isolated from mixed stage cultures of *C.elegans* (**Lane 1**) and from *C.elegans* embryos (**Lane 2**). The denaturing agarose gel used to generate the Northern blots is shown in Figure 5.3A.

A: Northern blot probed with the cDNA clone of *cpr-4* (cm14e3)

B: Northern blot probed with the cDNA clone of *cpr-5* (cm04d10)

C: Northern blot probed with the cDNA clone of *cpr-6* (cm01a5)

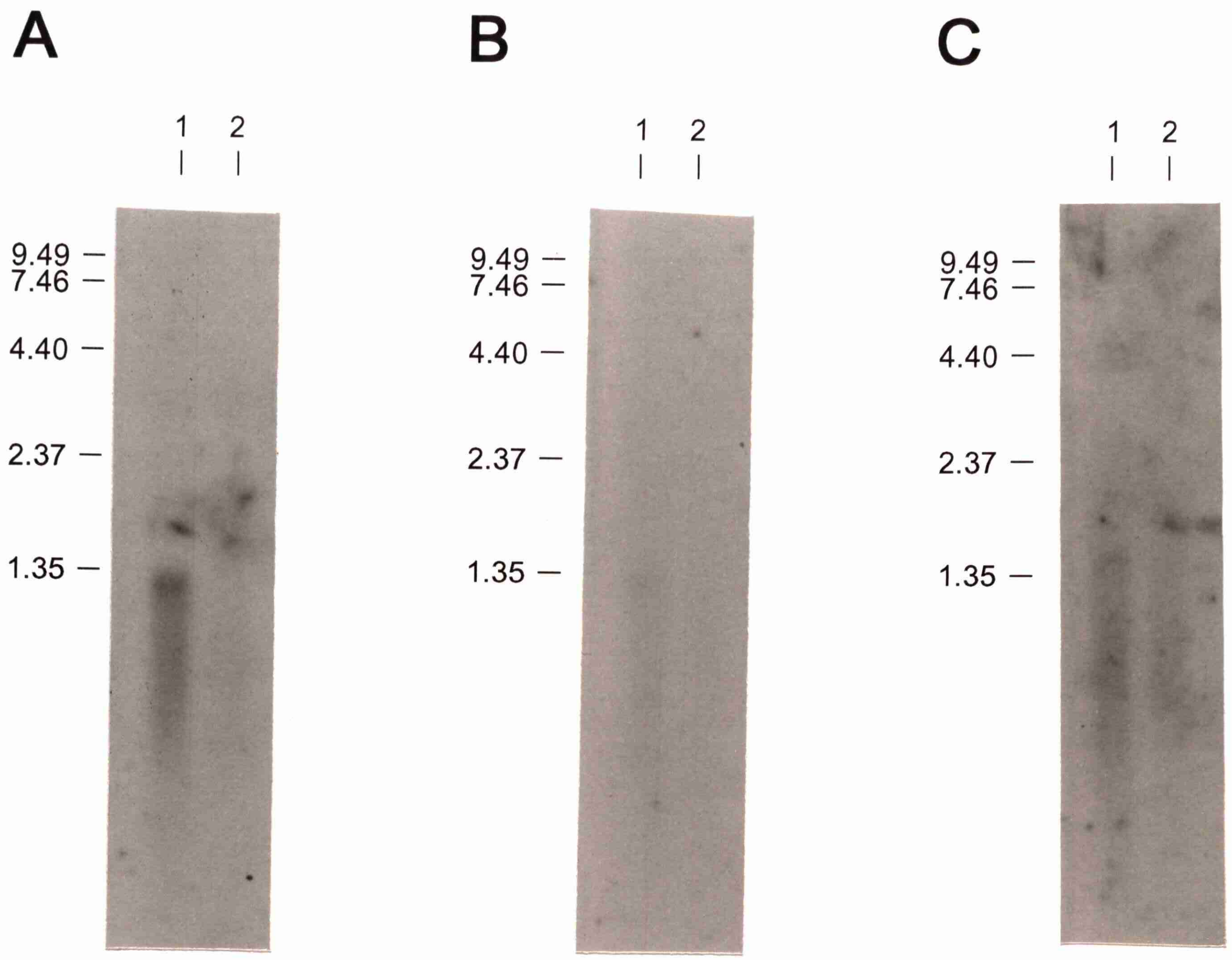
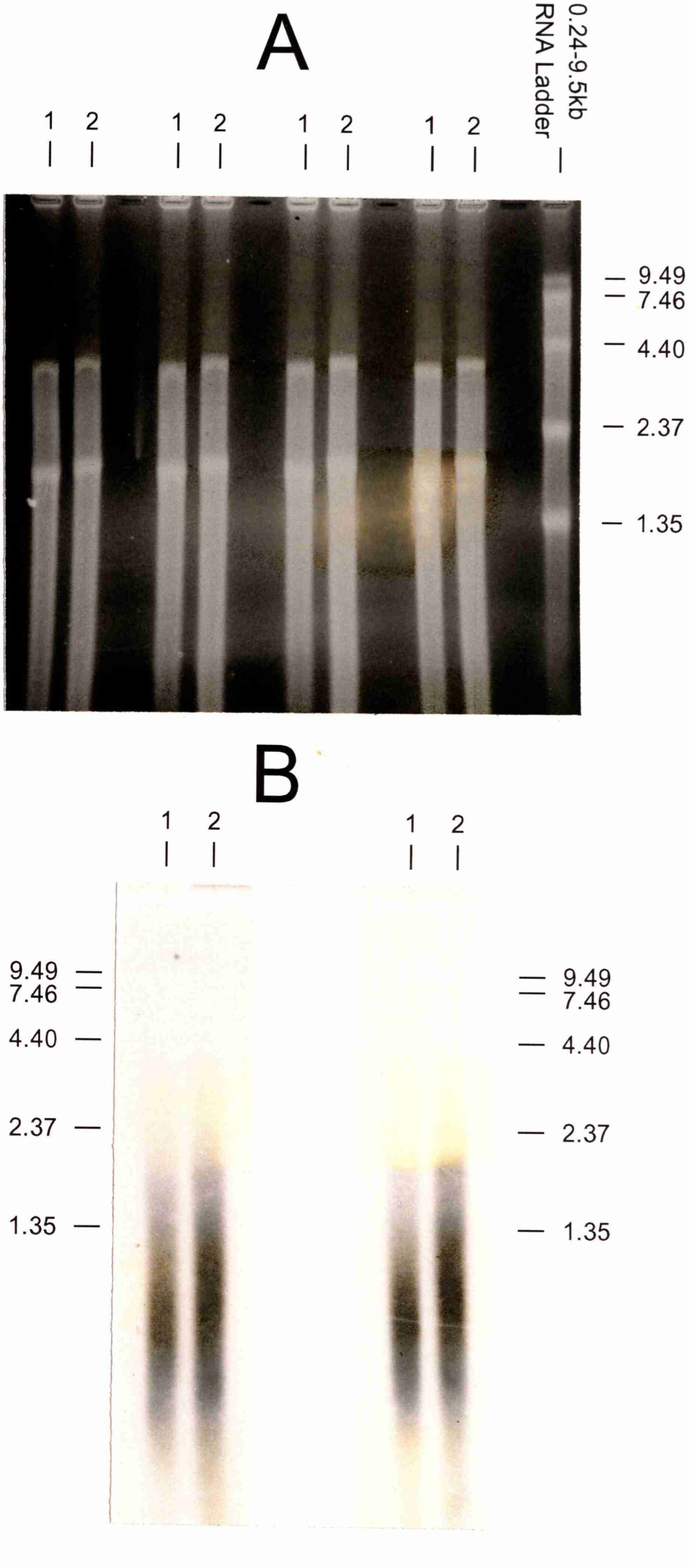

A
B
C
1
2
9.49
7.46
4.40
2.37
1.35

Figure 5.3

A: An ethidium bromide stained 1.0% denaturing agarose gel showing four aliquots of 4µg of total RNA isolated from mixed stage cultures of *C.elegans* **(Lane 1)** and *C.elegans* embryos **(Lane 2)** prior to Northern blotting.

B: The autoradiographic results obtained after high stringency washing of two Northern blots hybridised with a 7kb *Bam*HI rDNA fragment containing a single repeat of the 26s and 18s ribosomal DNA sequences. The Northern blots were generated from the denaturing agarose gel shown in **A**.



A
1
2
1
2
1
2
1
2
0.24-9.5kb RNA Ladder
9.49
7.46
4.40
2.37
1.35
B
1
2
1
2
9.49
7.46
4.40
2.37
1.35
9.49
7.46
4.40
2.37
1.35

Figure 5.4

The autoradiographic results obtained from Northern blot analysis of poly(A)$^+$ selected RNA isolated from mixed stage cultures of *C.elegans* **(Lane 1)** and *C.elegans* embryos **(Lane 2)**

The Northern blot was hybridised with a riboprobe generated from the *cpr-3* cDNA clone (cm12b6) and washed at high stringency.

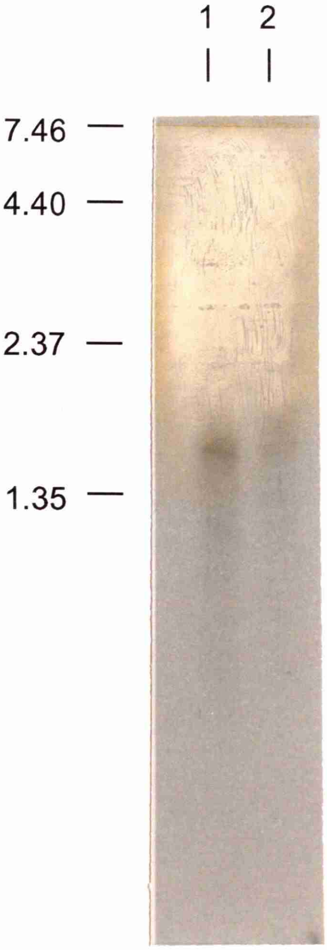

1
2
7.46
4.40
2.37
1.35

Figure 5.5

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-3*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-3* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5′ end-labelled with (γ^{32}P)ATP. The Southern blots were washed at medium stringency.

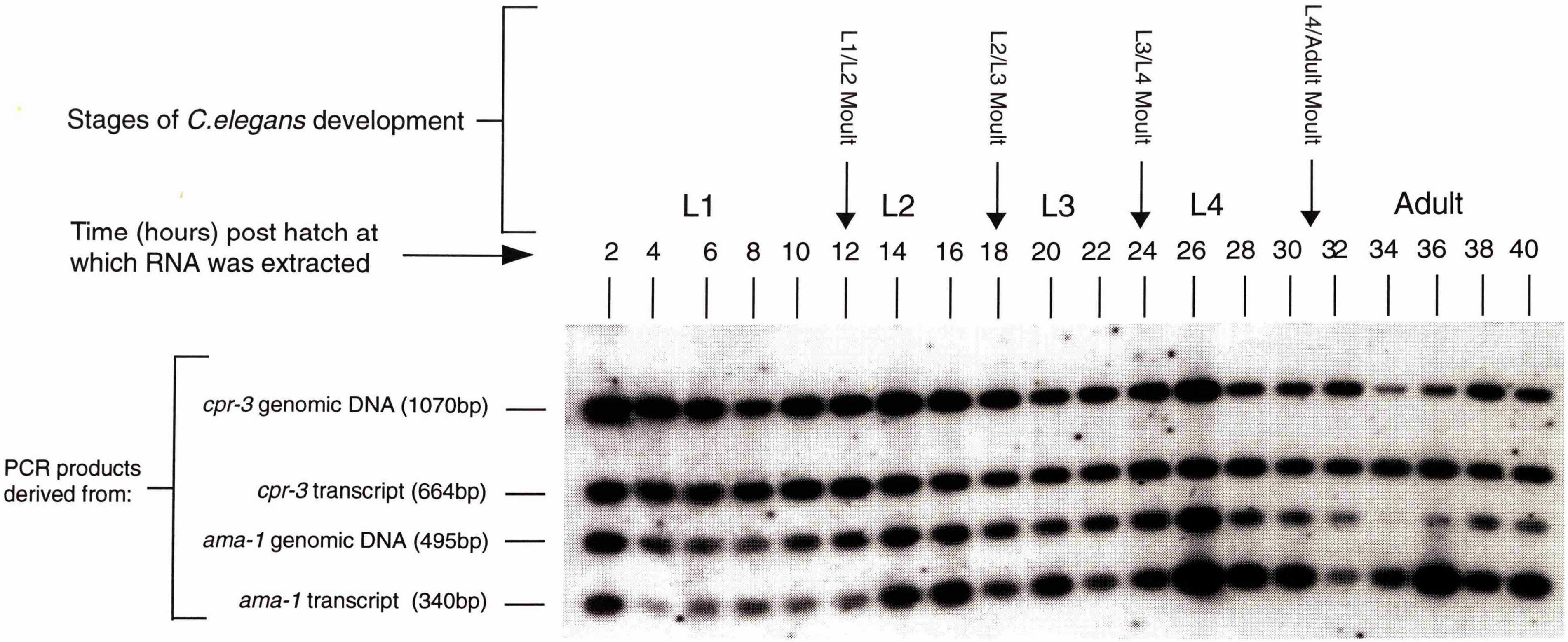

Stages of *C.elegans* development
L1/L2 Moult
L2/L3 Moult
L3/L4 Moult
L4/Adult Moult
L1
L2
L3
L4
Adult
Time (hours) post hatch at which RNA was extracted
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
PCR products derived from:
cpr-3 genomic DNA (1070bp)
cpr-3 transcript (664bp)
ama-1 genomic DNA (495bp)
ama-1 transcript (340bp)

Figure 5.6

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-4*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-4* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5′ end-labelled with (γ^{32}P)ATP. The Southern blots were washed at medium stringency.

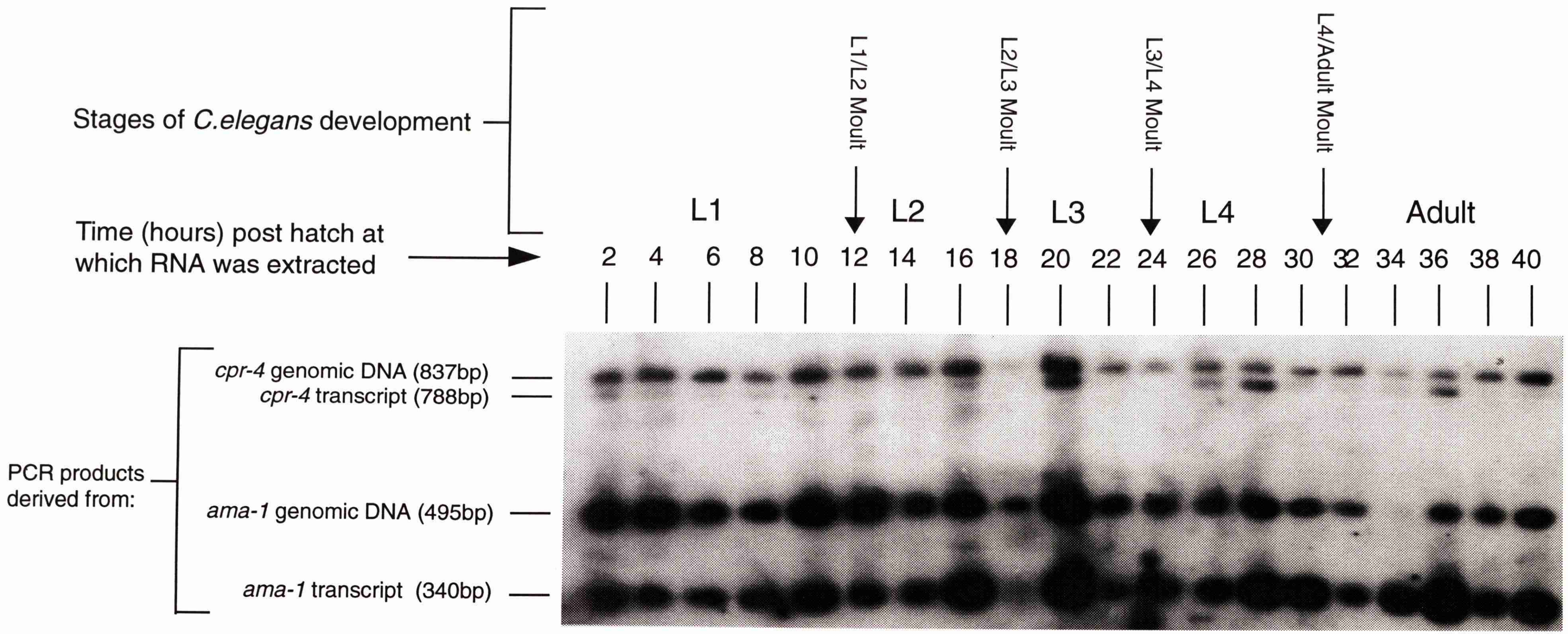

Stages of *C.elegans* development
Time (hours) post hatch at which RNA was extracted
L1
L1/L2 Moult
L2
L2/L3 Moult
L3
L3/L4 Moult
L4
L4/Adult Moult
Adult
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
PCR products derived from:
cpr-4 genomic DNA (837bp)
cpr-4 transcript (788bp)
ama-1 genomic DNA (495bp)
ama-1 transcript (340bp)

Figure 5.7

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-5*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-5* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5′ end-labelled with (γ^{32}P)ATP. The Southern blots were washed at medium stringency.

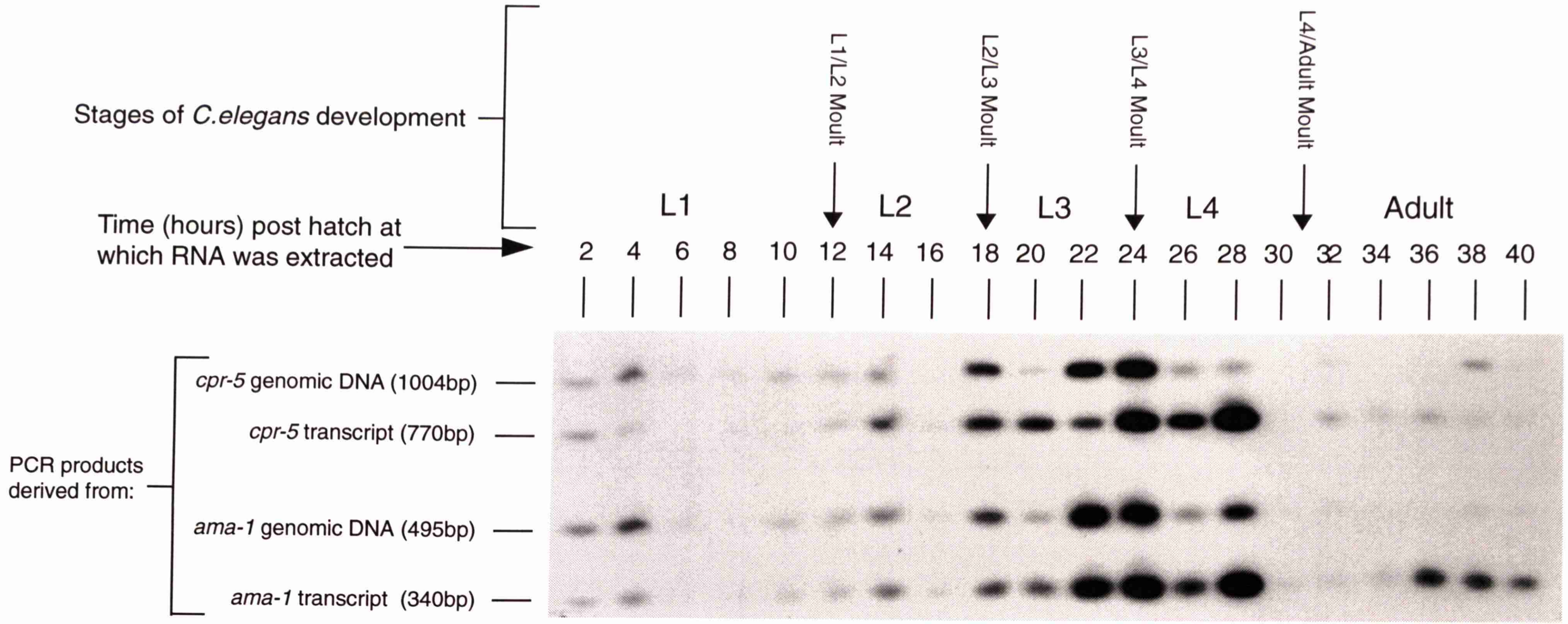

Stages of C.elegans development
L1
L1/L2 Moult
L2
L2/L3 Moult
L3
L3/L4 Moult
L4
L4/Adult Moult
Adult
Time (hours) post hatch at which RNA was extracted
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
PCR products derived from:
cpr-5 genomic DNA (1004bp)
cpr-5 transcript (770bp)
ama-1 genomic DNA (495bp)
ama-1 transcript (340bp)

Figure 5.8

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-6*

Each of 20 cDNA samples, generated from total RNA isolated at two hour intervals post hatch, was amplified by PCR. Each PCR reaction was performed using primers specific for *cpr-6* and the internal control gene, *ama-1*. The PCR reactions were electrophoresed through a 1.2% agarose gel, Southern blotted to Hybond N (Amersham International plc) and hybridised with the same primers used in the PCR reaction, 5′ end-labelled with (γ^{32}P)ATP. The Southern blots were washed at medium stringency.

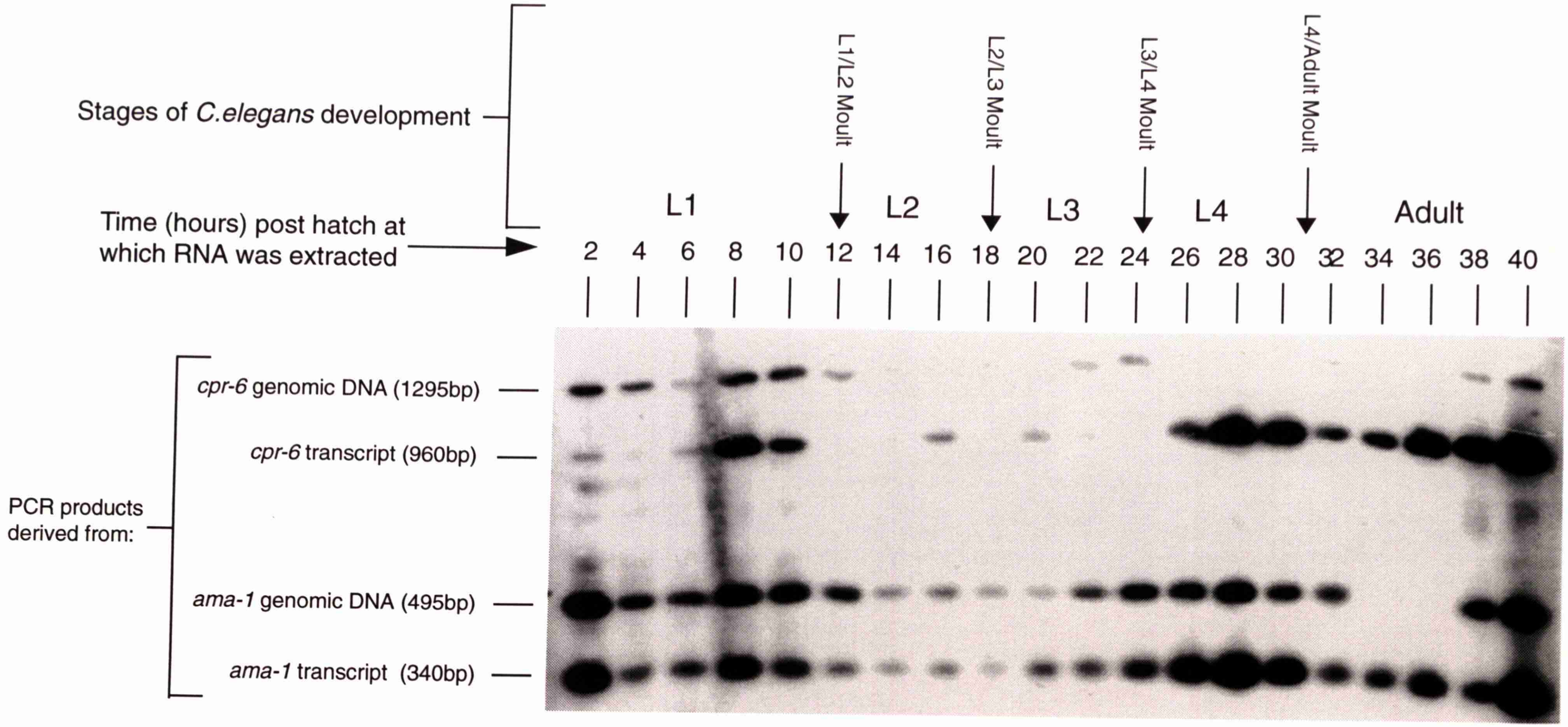

Stages of *C.elegans* development
Time (hours) post hatch at which RNA was extracted
L1
L1/L2 Moult
L2
L2/L3 Moult
L3
L3/L4 Moult
L4
L4/Adult Moult
Adult
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
PCR products derived from:
cpr-6 genomic DNA (1295bp)
cpr-6 transcript (960bp)
ama-1 genomic DNA (495bp)
ama-1 transcript (340bp)

Figure 5.9

Autoradiographic results obtained from s-q rtPCR analysis of *dpy-7*

The autoradiographic results of s-q rtPCR analysis of *dpy-7* expression were obtained by Dr.I.L.Johnstone who also donated the autoradiograph shown in this figure.

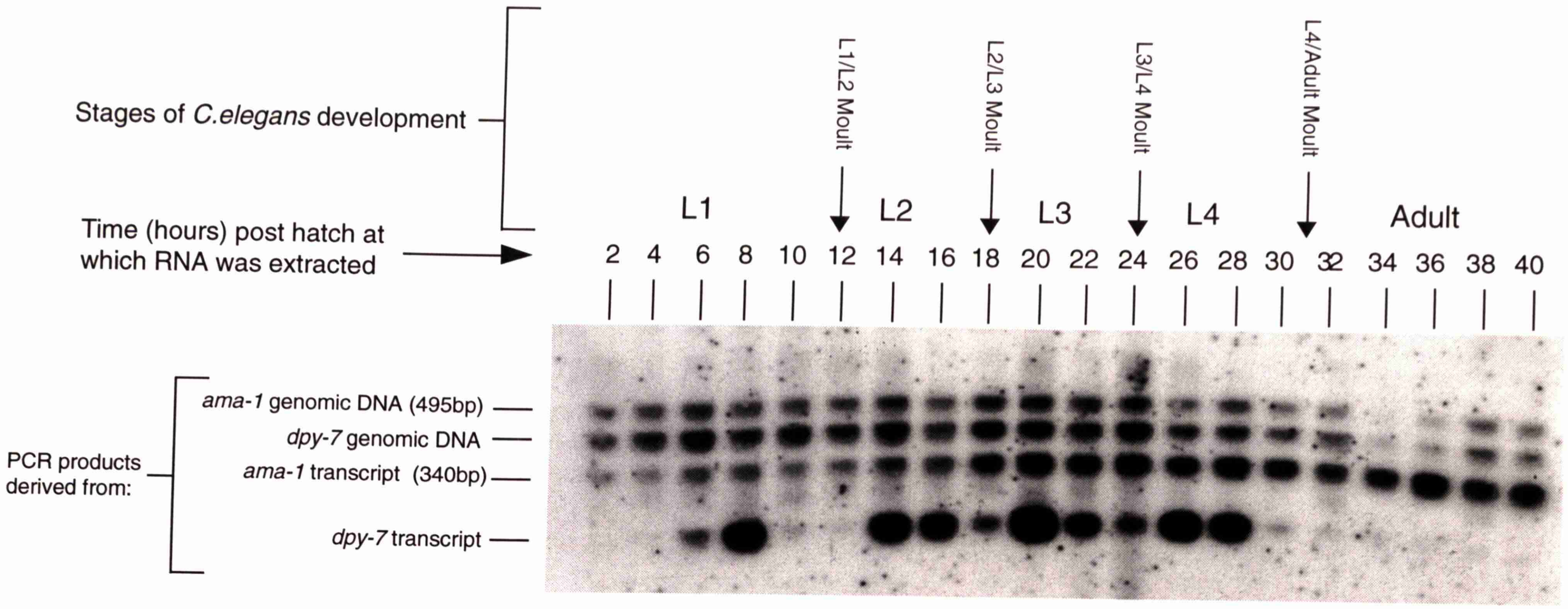

Stages of *C.elegans* development
Time (hours) post hatch at which RNA was extracted
L1
L1/L2 Moult
L2
L2/L3 Moult
L3
L3/L4 Moult
L4
L4/Adult Moult
Adult
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
PCR products derived from:
ama-1 genomic DNA (495bp)
dpy-7 genomic DNA
ama-1 transcript (340bp)
dpy-7 transcript

Figure 5.10

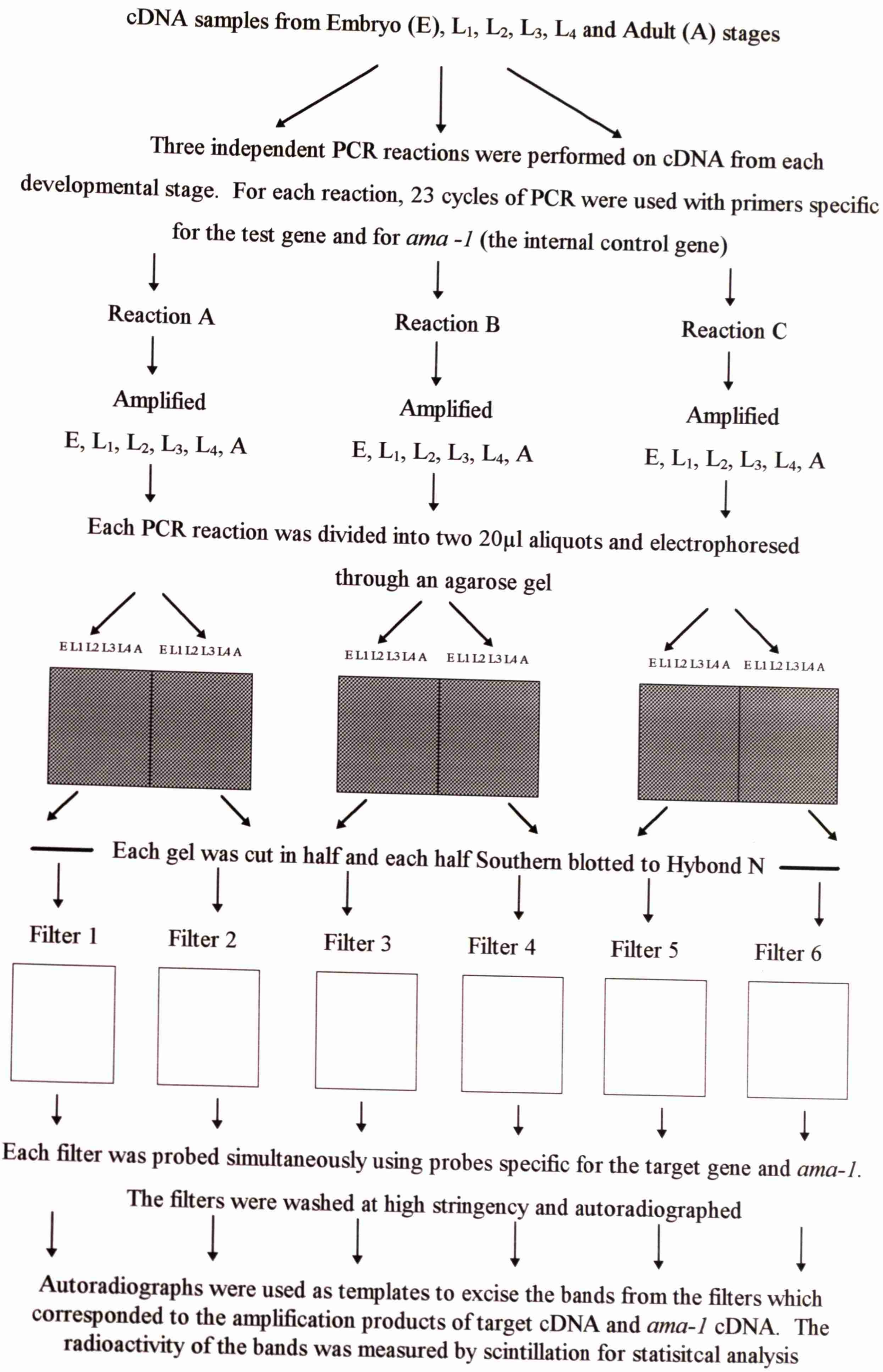
A summary of the modified s-q rtPCR approach used to analyse the temporal patterns of expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

Notes: Generation of the cDNA samples:

The Embryo cDNA was generated from RNA extracted from mixed stage embryos. The L_1, L_2, L_3, L_4 and Adult cDNA samples were generated by pooling the appropriate samples from the 20 staged cDNA samples generated from total RNA isolated at two hour intervals post embryo hatch.

The samples were pooled in the following manner:

L_1:	Samples 1-6	(from 2-12 hours post hatch)
L_2:	Samples 7-9	(from 12-18 hours post hatch)
L_3:	Samples 10-12	(from 18-24 hours post hatch)
L_4:	Samples 13-15	(from 24-30 hours post hatch)
Adult:	Samples 16-20	(from 30-40 hours post hatch)



cDNA samples from Embryo (E), L_1, L_2, L_3, L_4 and Adult (A) stages

Three independent PCR reactions were performed on cDNA from each developmental stage. For each reaction, 23 cycles of PCR were used with primers specific for the test gene and for *ama -1* (the internal control gene)

Reaction A

Reaction B

Reaction C

Amplified

E, L_1, L_2, L_3, L_4, A

Each PCR reaction was divided into two 20µl aliquots and electrophoresed through an agarose gel

Each gel was cut in half and each half Southern blotted to Hybond N

Each filter was probed simultaneously using probes specific for the target gene and *ama-1*. The filters were washed at high stringency and autoradiographed

Autoradiographs were used as templates to excise the bands from the filters which corresponded to the amplification products of target cDNA and *ama-1* cDNA. The radioactivity of the bands was measured by scintillation for statisitcal analysis

Figure 5.11

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-3* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction **(Reaction A)**.

Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction **(Reaction B)**.

Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction **(Reaction C)**.

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

Contaminating genomic DNA amplified with the *cpr-3* primers:	1070bp
cDNA derived from *cpr-3* transcript amplified with the *cpr-3* primers:	664bp
Contaminating genomic DNA amplified with the *ama-1* primers:	495bp
cDNA derived from *ama-1* transcript amplified with the *ama-1* primers:	340bp

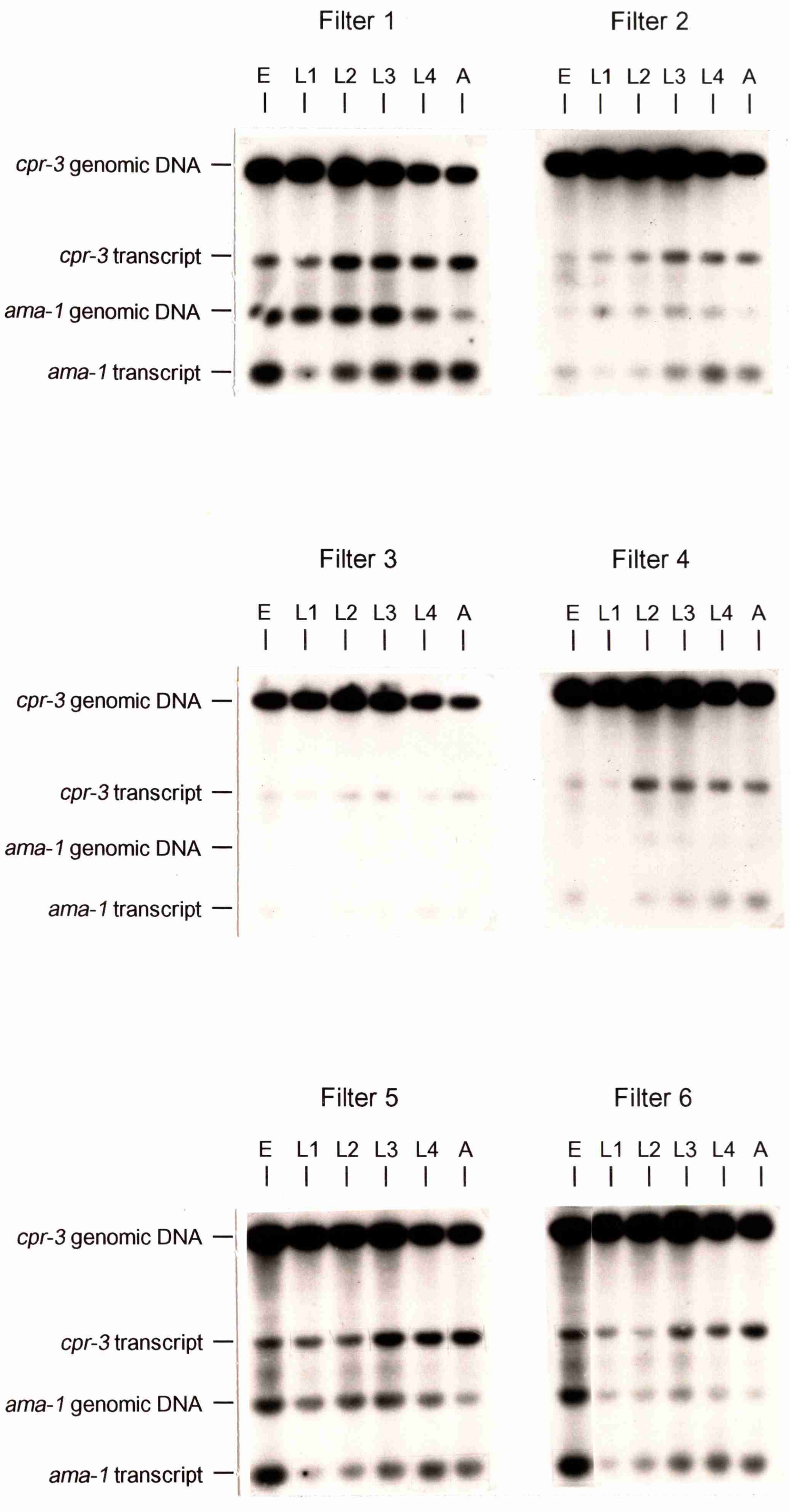

Filter 1
Filter 2
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-3 genomic DNA
cpr-3 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 3
Filter 4
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-3 genomic DNA
cpr-3 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 5
Filter 6
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-3 genomic DNA
cpr-3 transcript
ama-1 genomic DNA
ama-1 transcript

Figure 5.12

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-4* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction **(Reaction A)**.
Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction **(Reaction B)**.
Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction **(Reaction C)**.

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

cDNA derived from *cpr-4* transcript amplified with the *cpr-4* primers:	418bp
Contaminating genomic DNA amplified with the *ama-1* primers:	495bp
cDNA derived from *ama-1* transcript amplified with the *ama-1* primers:	340bp

Note: Filters 3 and 6 were photographed in the reverse orientation.

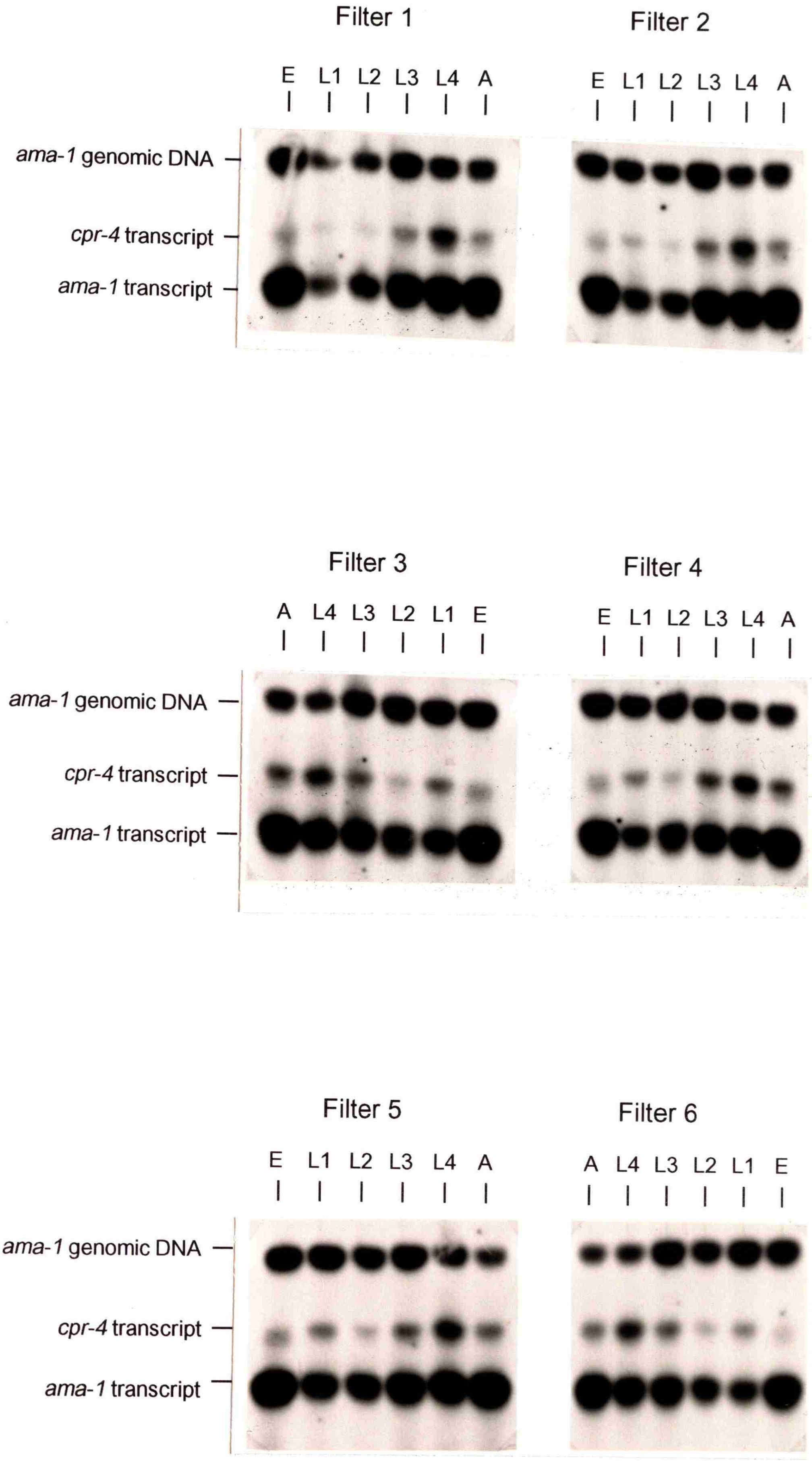

Filter 1
E
L1
L2
L3
L4
A
Filter 2
E
L1
L2
L3
L4
A
ama-1 genomic DNA
cpr-4 transcript
ama-1 transcript
Filter 3
A
L4
L3
L2
L1
E
Filter 4
E
L1
L2
L3
L4
A
ama-1 genomic DNA
cpr-4 transcript
ama-1 transcript
Filter 5
E
L1
L2
L3
L4
A
Filter 6
A
L4
L3
L2
L1
E
ama-1 genomic DNA
cpr-4 transcript
ama-1 transcript

Figure 5.13

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-5* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction **(Reaction A)**.

Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction **(Reaction B)**.

Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction **(Reaction C)**.

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

Contaminating genomic DNA amplified with the *cpr-5* primers:	1004bp
cDNA derived from *cpr-5* transcript amplified with the *cpr-5* primers:	770bp
Contaminating genomic DNA amplified with the *ama-1* primers:	495bp
cDNA derived from *ama-1* transcript amplified with the *ama-1* primers:	340bp

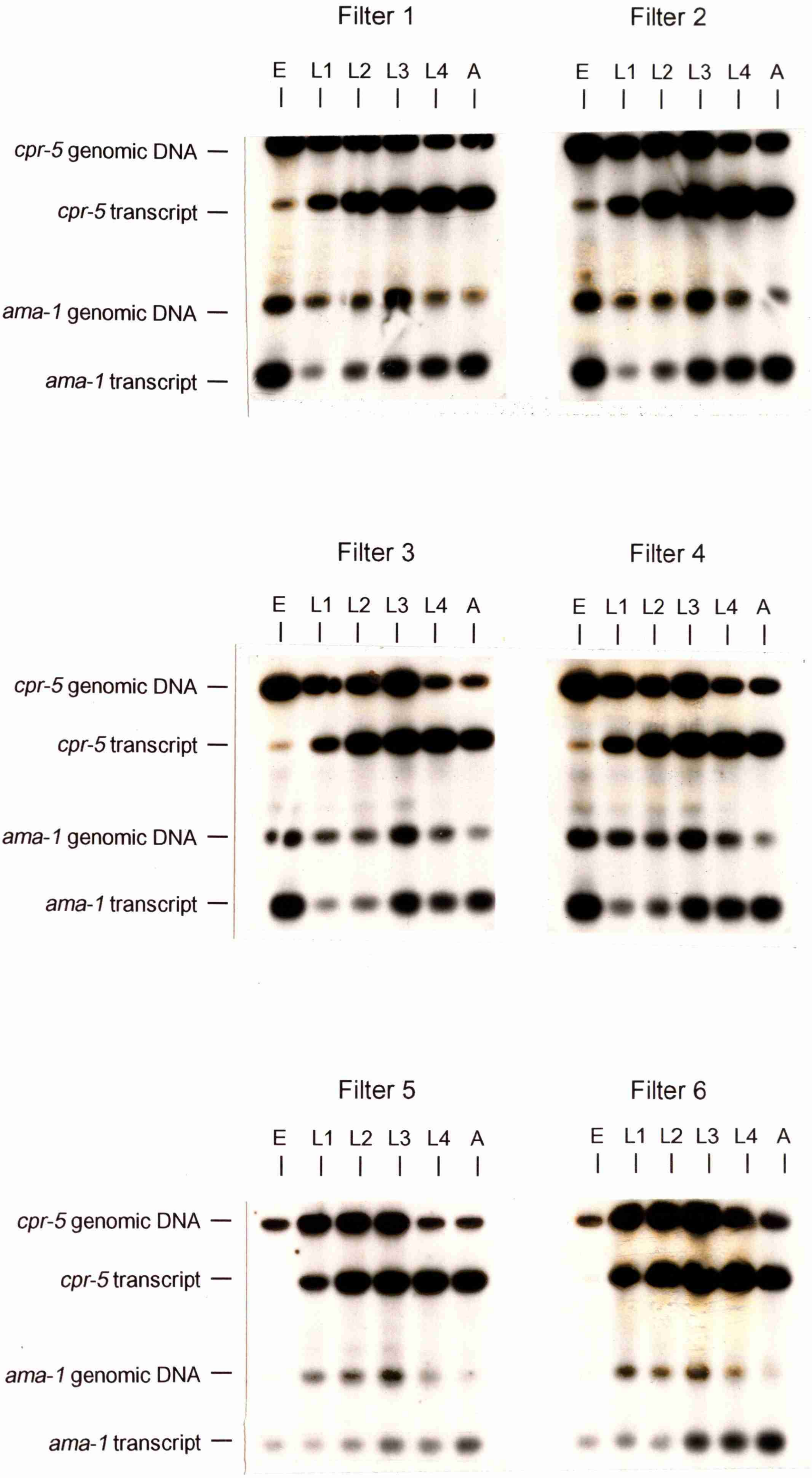

Filter 1
E L1 L2 L3 L4 A
Filter 2
E L1 L2 L3 L4 A
cpr-5 genomic DNA
cpr-5 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 3
E L1 L2 L3 L4 A
Filter 4
E L1 L2 L3 L4 A
cpr-5 genomic DNA
cpr-5 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 5
E L1 L2 L3 L4 A
Filter 6
E L1 L2 L3 L4 A
cpr-5 genomic DNA
cpr-5 transcript
ama-1 genomic DNA
ama-1 transcript

Figure 5.14

Autoradiographic results obtained from s-q rtPCR analysis of *cpr-6* using the modified protocol

A summary of the protocol is shown in Figure 5.10.

Filters 1 and 2 were obtained from independent Southern blots of the same PCR reaction (**Reaction A**).
Filters 3 and 4 were obtained from independent Southern blots of the same PCR reaction (**Reaction B**).
Filters 5 and 6 were obtained from independent Southern blots of the same PCR reaction (**Reaction C**).

The template from which each amplification product was derived is indicated.

The sizes of the amplification products are as follows:

Contaminating genomic DNA amplified with the *cpr-6* primers:	1295bp
cDNA derived from *cpr-6* transcript amplified with the *cpr-6* primers:	960bp
Contaminating genomic DNA amplified with the *ama-1* primers:	495bp
cDNA derived from *ama-1* transcript amplified with the *ama-1* primers:	340bp

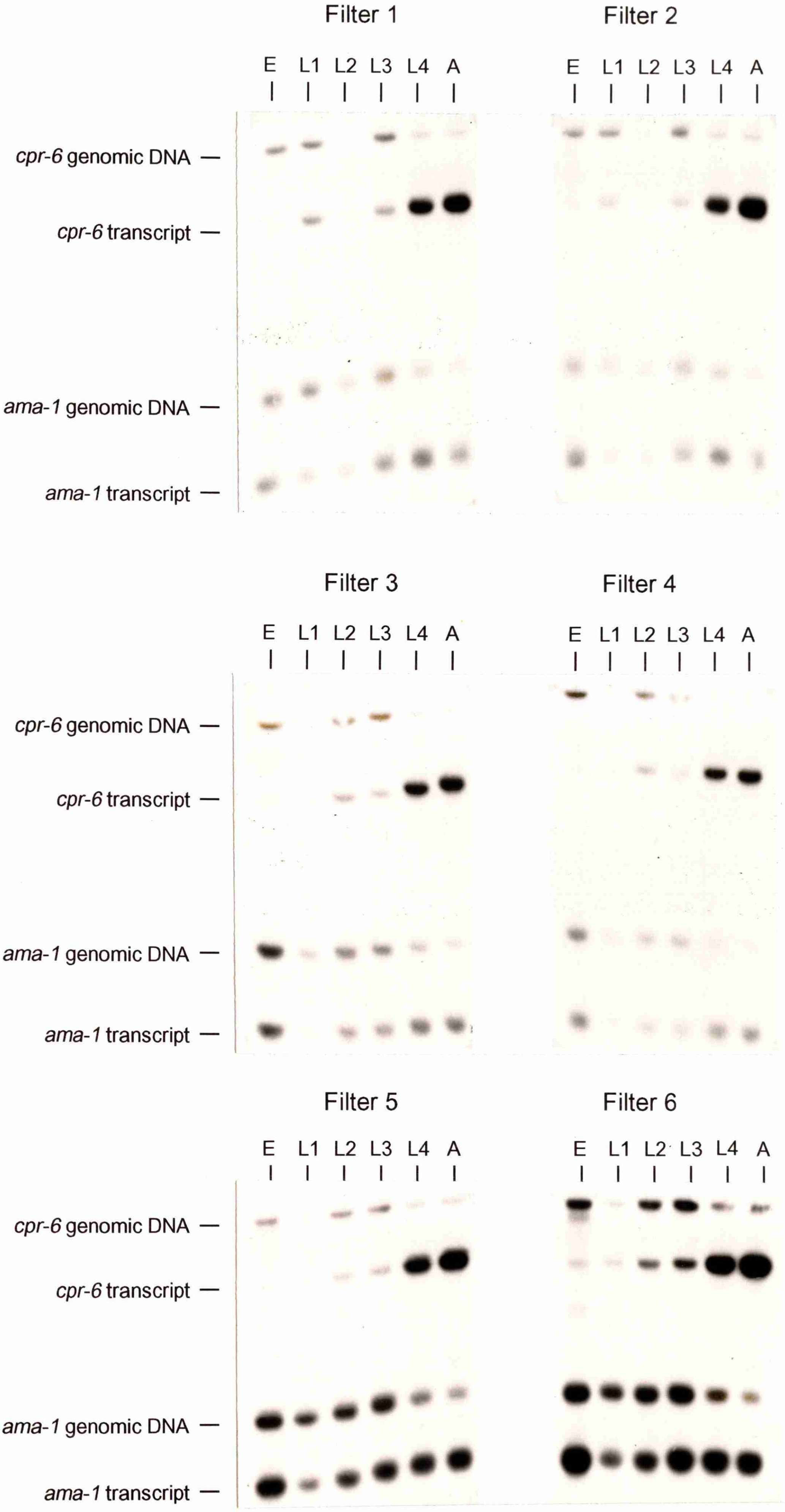

Filter 1
Filter 2
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-6 genomic DNA
cpr-6 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 3
Filter 4
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-6 genomic DNA
cpr-6 transcript
ama-1 genomic DNA
ama-1 transcript
Filter 5
Filter 6
E L1 L2 L3 L4 A
E L1 L2 L3 L4 A
cpr-6 genomic DNA
cpr-6 transcript
ama-1 genomic DNA
ama-1 transcript

Figure 5.15

Graphical representations of the relative transcript abundance ratios obtained for *cpr-3* (graphs **A1** and **A2**), *cpr-4* (graphs **B1** and **B2**), *cpr-5* (graphs **C1** and **C2**) and *cpr-6* (graphs **D1** and **D2**) using s-q rtPCR

A summary of the protocol is shown in Figure 5.10.

Each graph shows the relative transcript abundance ratio (Y-axis) for each developmental stage (X-axis). Each pair of graphs was generated from the radioactivity data obtained for *cpr-3*, cpr-*4*, *cpr-5* and *cpr-6*, shown in Tables 5.1, 5.2, 5.3 and 5.4 respectively. Graphs **A1**, **B1**, **C1** and **D1** show the relative transcript abundance ratios for *cpr-3*, cpr-*4*, *cpr-5* and *cpr-6* respectively, obtained from filters 1, 3 and 5 for each gene*. Graphs **A2**, **B2**, **C2** and **D2** show the relative transcript abundance ratios for *cpr-3*, cpr-*4*, *cpr-5* and *cpr-6* respectively, obtained from filters 2, 4 and 6 for each gene. Each pair of graphs shows data generated from the three independent sets of PCR reactions (A, B and C, Figure 5.10) performed for each gene.

KEY:

Relative transcript abundance ratios determined from PCR reaction A (filters 1 and 2, for each gene): ————————

Relative transcript abundance ratios determined from PCR reaction B (filters 3 and 4, for each gene): - - - - - - - - - - -

Relative transcript abundance ratios determined from PCR reaction C (filters 5 and 6, for each gene): – · – · – · –

* For *cpr-6*, the radioactivity data from filters 3 and 5 were lost and therefore only the relative transcript abundance ratios obtained from filter 1 are shown.

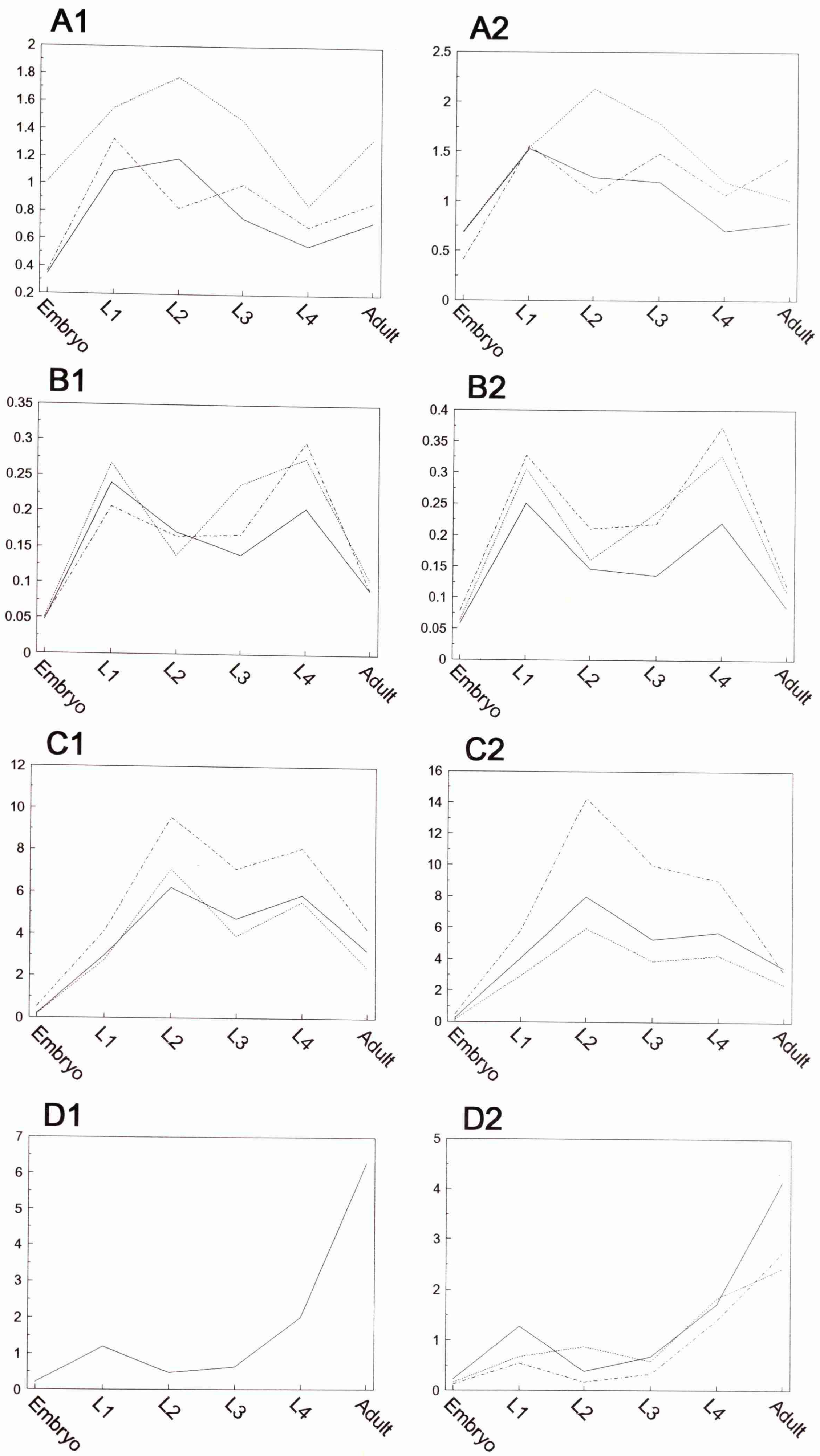

A1
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
Embryo
L1
L2
L3
L4
Adult
A2
2.5
2
1.5
1
0.5
0
Embryo
L1
L2
L3
L4
Adult
B1
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Embryo
L1
L2
L3
L4
Adult
B2
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Embryo
L1
L2
L3
L4
Adult
C1
12
10
8
6
4
2
0
Embryo
L1
L2
L3
L4
Adult
C2
16
14
12
10
8
6
4
2
0
Embryo
L1
L2
L3
L4
Adult
D1
7
6
5
4
3
2
1
0
Embryo
L1
L2
L3
L4
Adult
D2
5
4
3
2
1
0
Embryo
L1
L2
L3
L4
Adult

Figure 5.16

The temporal patterns of expression of *cpr -3* (graph **A**), *cpr-4* (graph **B**), *cpr-5* (graph **C**) and *cpr-6* (graph **D**) generated by s-q rtPCR

A summary of the protocol is shown in Figure 5.10. Each graph shows the mean and standard deviation of the relative transcript abundance ratio (Y-axis) for each developmental stage (X-axis), determined from three independent PCR reactions. The radioactivity data used to generate graphs **A**, **B**, **C** and **D** are shown in Tables 4.1, 4.2, 4.3 and 4.4 respectively. For graphs **A**, **B** and **C**, the mean and standard deviation of the relative transcript abundance ratio for each developmental stage was calculated from radioactivity data obtained from filters 1, 3 and 5. For graph **D**, the mean and standard deviation of the relative transcript abundance ratio for each developmental stage was calculated from radioactivity data obtained from filters 2, 4 and 6.

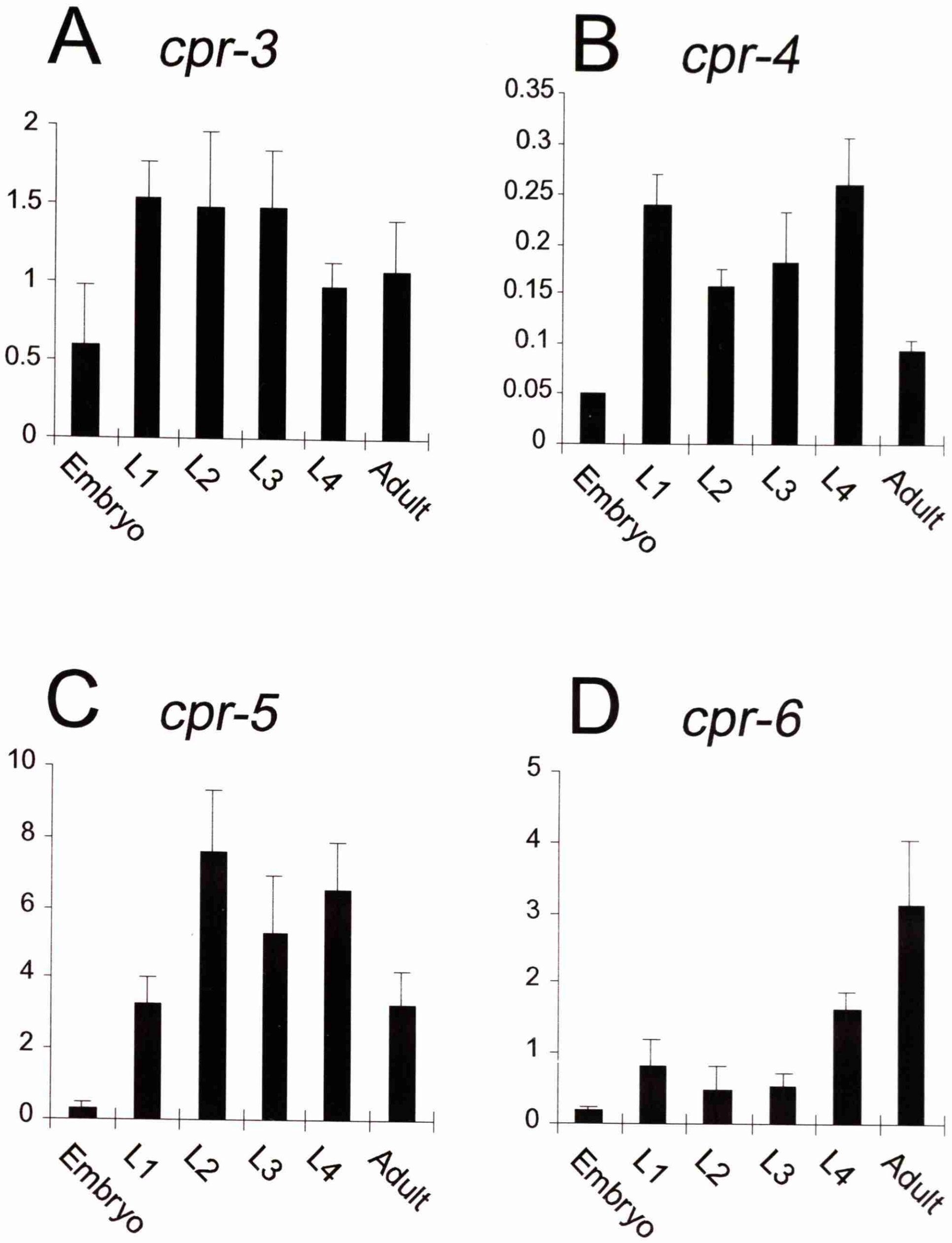

A
cpr-3
2
1.5
1
0.5
0
Embryo
L1
L2
L3
L4
Adult
B
cpr-4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Embryo
L1
L2
L3
L4
Adult
C
cpr-5
10
8
6
4
2
0
Embryo
L1
L2
L3
L4
Adult
D
cpr-6
5
4
3
2
1
0
Embryo
L1
L2
L3
L4
Adult

Figure 5.17

A schematic diagram showing the features of the pPD21.28 and pPD95.03 *lacZ* fusion vectors used to study the spatial expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The figure shows the relevant features of the pPD21.28 vector. The pPD95.03 vector was derived from pPD21.28 and therefore only the additional features of this vector are shown.

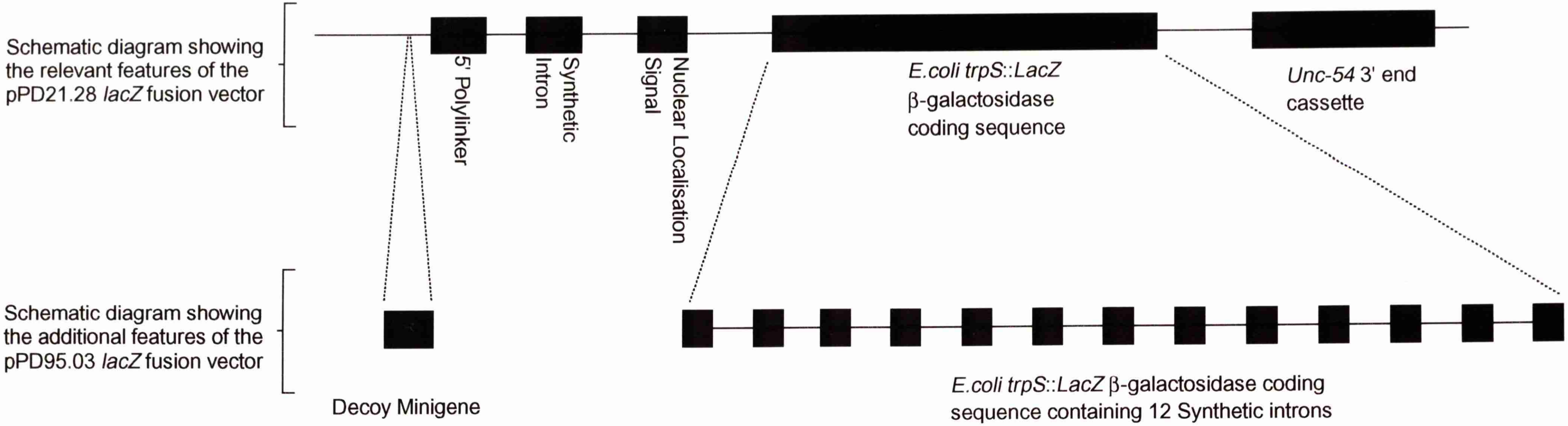
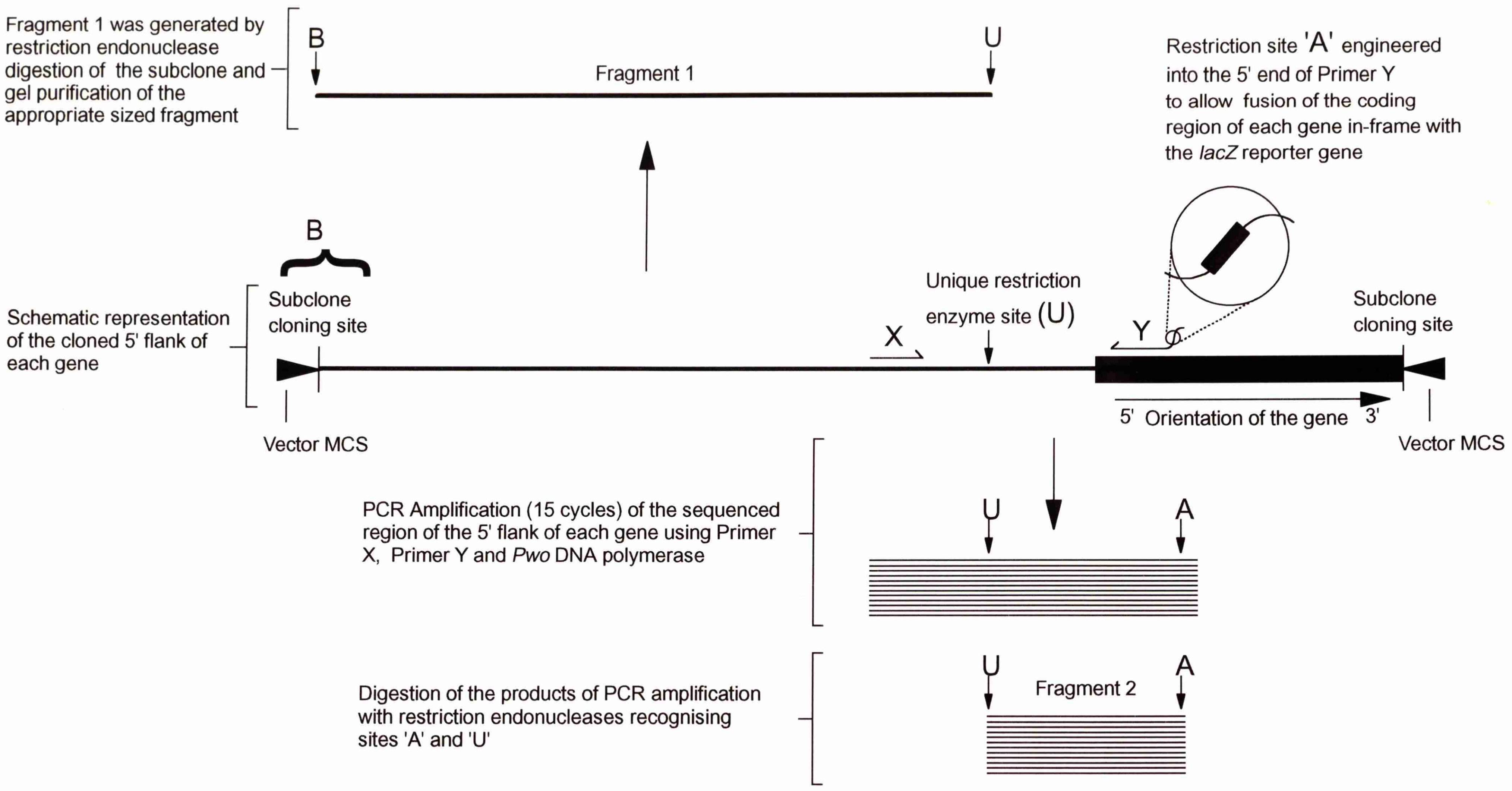

Schematic diagram showing the relevant features of the pPD21.28 *lacZ* fusion vector
5' Polylinker
Synthetic Intron
Nuclear Localisation Signal
E.coli trpS::*LacZ* β-galactosidase coding sequence
Unc-54 3' end cassette
Schematic diagram showing the additional features of the pPD95.03 *lacZ* fusion vector
Decoy Minigene
E.coli trpS::*LacZ* β-galactosidase coding sequence containing 12 Synthetic introns

Figure 5.18

A schematic diagram showing the approach used to obtain the fragments necessary to generate in-frame translational fusions of the 5′ flank regions of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. For each gene, Fragment 1 was prepared by digestion of the appropriate subclone with restriction endonucleases recognising sites 'B' and 'U' and gel purification of the appropriate fragment. Fragment 2 was generated by 15 cycles of PCR using, primers 'X' and 'Y', *Pwo* DNA polymerase and the appropriate subclone as template. The PCR products were gel purified, digested with restriction endonucleases recognising sites 'A' and 'U' and subcloned either into pBluescript II SK- or directly into the appropriate *lacZ* fusion vector. The restriction endonucleases recognising sites 'A', 'B' and 'U' are given in Table 5.6A. The sizes of Fragment 1 and Fragment 2 for each gene are given in Table 5.6B.

The subclones used for the preparation of Fragment 1 and Fragment 2 for each gene are as follows:

cpr-3: CL#034/1.7/KS+
cpr-4: ZK1055/2.8/KS+
cpr-5: W02B2/3/KS+
cpr-6: C25B8/3.8/KS+



Fragment 1 was generated by restriction endonuclease digestion of the subclone and gel purification of the appropriate sized fragment

Restriction site 'A' engineered into the 5' end of Primer Y to allow fusion of the coding region of each gene in-frame with the *lacZ* reporter gene

Schematic representation of the cloned 5' flank of each gene

PCR Amplification (15 cycles) of the sequenced region of the 5' flank of each gene using Primer X, Primer Y and *Pwo* DNA polymerase

Digestion of the products of PCR amplification with restriction endonucleases recognising sites 'A' and 'U'

Figure 5.19

The sequenced regions of the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The unique restriction site 'U' for each gene is boxed and the putative translation initiation codon is underlined.

cpr-3 5' Flank

```
                                        -320 tccagacaatttaccctaaa -301
-300 gtgtgaccaaatgcctacattacacacacctcaacgctttgataagcctgataagctccc -241
-240 tcctaattcataatgatgatatcaacgaaggtgataattgatttcttgattcggtgacac -181
-180 taacctccgtgttctcgactgatgactaactttttttttgaatagacaaacataagaggg -121
-120 gaaaatgtcgaaattttagtgataagggaaagctacaaaatgtgcggtgtgtgtttgctg -61
 -60 aaacgcttctttccagttttcagttttccctcaaatttcaaaaattgaatactaaagagaATG +3
```

The unique *Eco*RV site for *cpr-3* is boxed

cpr-4 5' Flank

```
-359  tagtcctcatagttcaaaccttcttgctacttttacacctaacctaaaaattgagtact -301
-300 tctaatctggttccaaatgataactttcgttgaaccacacaacttccaaactcttatcaa -241
-240 agttgcacgagatcattgtgctcaaatgatggtgctgcgtcacatgactacctcctaatt -181
-180 aggcattgtctatcgaaatttgcgctgccaggtaccgcaaatttttcaattttaatccc -121
-120 ggacgccaatgataagagataacgagcactcccgaactgataacgagtcaactataaaag -61
 -60 accatcgcaatgaagtaacttcagcatttgctctatcttgctatttgctcttttacaaaaATG +3
```

The unique *Kpn*I site for *cpr-4* is boxed

cpr-5 5' Flank

```
                                    -324 cgcttatgcttatgcttatgcttt -301
-300 tgcttatgcttatgcttatgcttatgcttaggctcaggcttaggcttaggcattaagctt -241
-240 aggtattaggctcatccctaattcctaatccccttagaattttacagtattaaaaaatg -181
-180 tttacatcaatgatctattccgacaagtgaaagtatatttggcgcgcgactcatttacgt -121
-120 cacattcctttaatttttaattcttaaaaaagatactgataaaattaactgataagaat -61
 -60 gatgcgtaccgactacttaatgagcattagacgcgaaccttcgcagacacttctctcataATG +3
```

The unique *Bss*HII site for *cpr-5* is boxed

cpr-6 5' Flank

```
                                        -320 attaggtttttccatcatat -301
-300 aaccctttcaaacgaaattaatgtgctaaatctgttaagtttcaatattttccttgtctt -241
-240 taggtcaatcttctttgccacacagttcaaactactaccgccgagtcacgtcacaccatc -181
-180 acaggatagtgaccggtcctaggatgtaccctgacactgtgatggacgcagccgacactc -121
-120 ttatcgaaatgcacagggccaaatttgataacgaaaacatgttctataaaagcatgctga -61
 -60 taaaagcgagcagtcaagcgacgacaacttgcgatcaacacgctgacgtcgacgccaacATG +3
```

The unique *Sal*I site for *cpr-6* is boxed

Figure 5.20

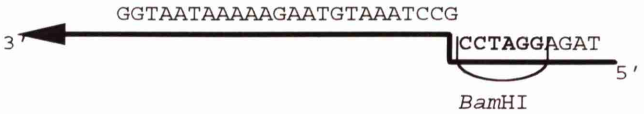
The region of each of the four genes, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* to which the 'Y' primers anneal

The putative ATG initiation codon of each gene is marked in bold. The restriction endonuclease sites 'A' are also indicated at the 5′ end of each primer.

cpr-3

ttcaaaaattgaatactaaagaga**ATG**CTGAAAGTGTACTTTTTGgtgagtcatcttcatcttcatttttccattatttttcttacatttagGCACTGTTTCTAGCCGGGTGCTCTGCATTTG

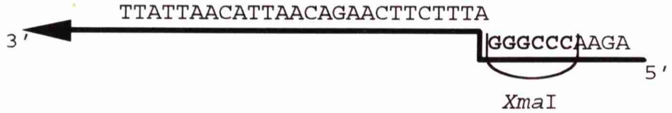
M L K V Y F L A L F L A G C S A F V



cpr-4

aaa**ATG**gtaagtggttgttttagcgtgaagtttaataattgtaattgtcttgaagAAATACCTCATTCTTGCTGCCTTGGTGGCTGTTACCGCCGGACTCGTTATTCCACTTGTTCCAAAAACC

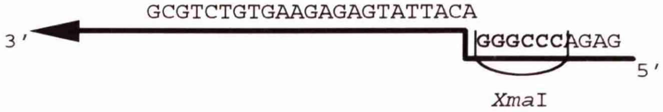
M K Y L I L A A L V A V T A G L V I P L V P K T



cpr-5

gatactgataaaattaactgataagaatgatgcgtaccgactacttaatgagcattagacgcgaaccttcgcagacacttctctcata**ATG**TGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCT

M W K L S A I L L V A A



cpr-6

aacacgctgaccgtcgacgccaac**ATG**gtaggcttttgaactttgaagtaatttttagagaaaatttgaattctagAAGACGTTGCTCTTCCTTTCCTGCATAGTGGTAGCAGCTTATTGCGCA

M K T L L F L S C I V V A A Y C A

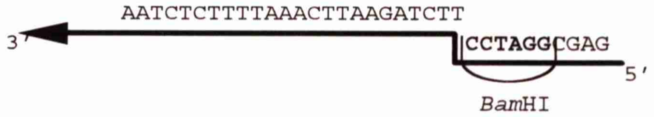


Figure 5.21

A schematic diagram summarising the approach used to insert the 5′ flanks of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* into the *lacZ* fusion vectors

'A', 'B' and 'U' represent the restriction endonuclease sites used to generate the *lacZ* fusion transgene constructs for each of the four genes. The actual restriction endonuclease sites used in the preparation of each construct are shown in Table 5.6A. For each gene, the 5′ flank was inserted into the *lacZ* fusion vectors as two fragments (Fragment 1 and Fragment 2). An outline of how Fragment 1 and Fragment 2 were generated for each gene is given in Figure 5.18 and the sizes of these fragments are given in Table 5.6B. The figure also indicates the names, orientation and approximate positions of the oligonucleotide primers used both for sequencing the *lacZ* fusion transgene constructs and for colony PCR.

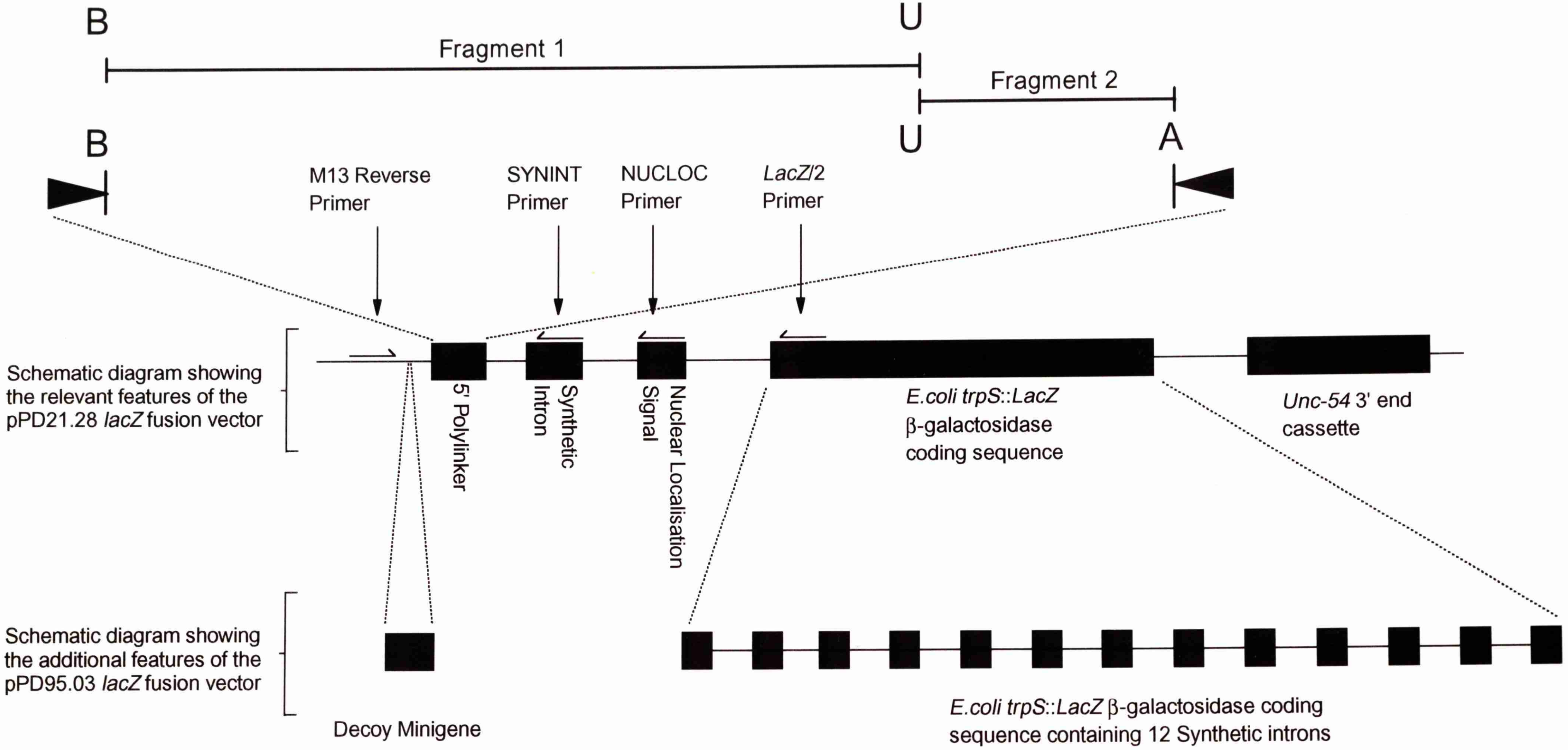

B
Fragment 1
U
Fragment 2
U
B
A
M13 Reverse Primer
SYNINT Primer
NUCLOC Primer
LacZ/2 Primer
Schematic diagram showing the relevant features of the pPD21.28 *lacZ* fusion vector
5' Polylinker
Synthetic Intron
Nuclear Localisation Signal
E.coli trpS::LacZ β-galactosidase coding sequence
Unc-54 3' end cassette
Schematic diagram showing the additional features of the pPD95.03 *lacZ* fusion vector
Decoy Minigene
E.coli trpS::LacZ β-galactosidase coding sequence containing 12 Synthetic introns

Figure 5.22

The DNA sequence of the junction between the 5′ flank of each gene and the appropriate *lacZ* fusion construct

Coding DNA sequence is shown in capitals while non-coding DNA sequence is shown in lowercase. The cloning site recognised by restriction endonuclease ‘A’ for each gene is marked in bold. Vertical bars indicate the open reading frame. The amino acid encoded by each codon is indicated below the appropriate codon.

A) *cpr-3*

Ai) Partial sequence of the 5’ flank of *cpr-3* including the putative ATG translation initiation codon and partial coding sequence

```
-20 aaaattgaatactaaagagaATGCTGAAAGTGTACTTTTTGgtgagtcatcttcatcttc +40
                         M  L  K  V  Y  F  L
+41 attttttccattatttttcttacatttagGCACTGTTTCTAGCCGGGTGCTCTGCATTTG +100
                                  A  L  F  L  A  G  C  S  A  F  V
```

Aii) Partial sequence of the 5’ flank of *cpr-3* fused to the polylinker of pPD95.03

```
-20 aaaattgaatactaaagagaATGCTGAAAGTGTACTTTTTGgtgagtcatcttcatcttc +40
                         M  L  K  V  Y  F  L
+41 attttttccattatttttcttacatttagGCGGATCCCCGGGATTGGCCAAAGGACCAA
                                  A  D  P  R  D  W  P  R  D  P
```

B) *cpr-4*

Bi) Partial sequence of the 5’ flank of *cpr-4* including the putative ATG translation initiation codon and partial coding sequence

```
-20 tatttgctcttttacaaaaATGgtaagtggttgttttagcgtgaagtttaataattgtaa +40
                        M
+41 ttgtcttgaagAAATACCTCATTCTTGCTGCCTTG +75
                K  Y  L  I  L  A  A  L
```

Bii) Partial sequence of the 5’ flank of *cpr-4* fused to the polylinker of pPD95.03

```
-20 tatttgctcttttacaaaaATGgtaagtggttgttttagcgtgaagtttaataattgtaa +40
                        M
+41 ttgtcttgaagAAATCCCGGGATTGGCCAAAG
                K  S  R  D  W  P  R
```

C) *cpr-5*

Ci) Partial sequence of the 5’ flank of *cpr-5* including the putative ATG translation initiation codon and partial coding sequence

```
-24 accttcgcagacacttctctcataATGTGGAAGCTCTCCGCTATTCTTCTCGTGGCTGCT +36
                             M  W  K  L  S  A  I  L  L  V  A  A
```

Cii) Partial sequence of the 5’ flank of *cpr-5* fused to the polylinker of pPD21.28

```
-24 accttcgcagacacttctctcataATGTCCCGGGATTGGCCAAAGGACCAA
                             M  S  R  D  W  P  R  D  P
```

D) *cpr-6*

Di) Partial sequence of the 5’ flank of *cpr-6* including the putative ATG translation initiation codon and partial coding sequence

```
-20 cgctgaccgtcgacgccaacATGgtaggcttttgaactttgaagtaatttttagagaaaa +40
                         M
+41 tttgaattctagAAGACGTTGCTCTTCCTTTCCTGC
                 K  T  L  L  F  L  S  C
```

Dii) Partial sequence of the 5’ flank of *cpr-6* fused to the polylinker of pPD95.03

```
-20 cgctgaccgtcgacgccaacATGgtaggcttttgaactttgaagtaatttttagagaaaa +40
                         M
+41 tttgaattctagAAGGATCCCCGGGATTGGCCAAAGGACCAA
                 K  D  P  R  D  W  P  R  D  P
```

Figure 5.23

Expression of the *cpr-3::lacZ* transgene fusion in transgenic *C.elegans*

Strains of *C.elegans* transgenic for the *cpr-3::lacZ* transgene fusion were fixed and stained for ß-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C.elegans* development.

A. Adult

B. L_2

C. L_1

D. Embryo (threefold stage)

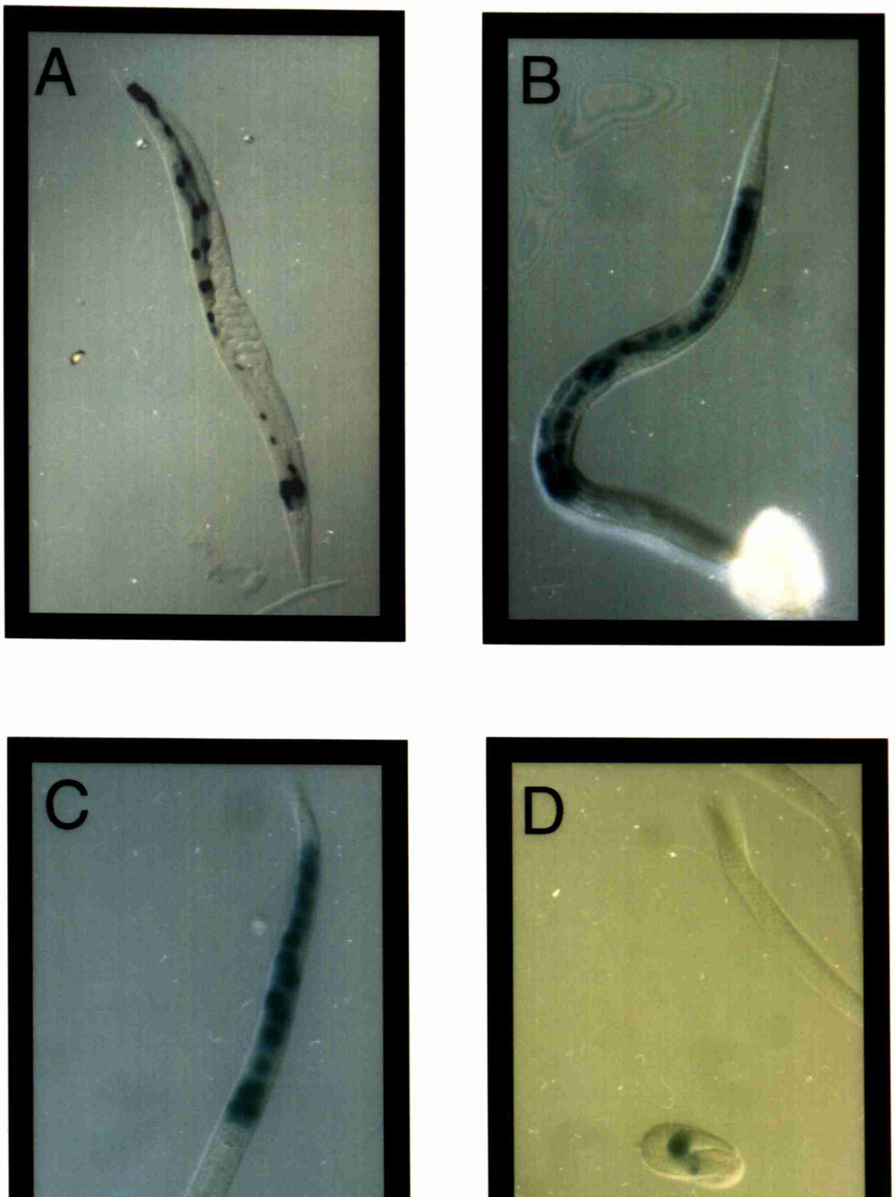

A
B
C
D

Figure 5.24

Expression of the *cpr-4::lacZ* transgene fusion in transgenic *C.elegans*

Strains of *C.elegans* transgenic for the *cpr-4::lacZ* transgene fusion were fixed and stained for ß-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C.elegans* development.

A. Adult

B. L_4

C. L_2

D. Embryo (threefold stage)

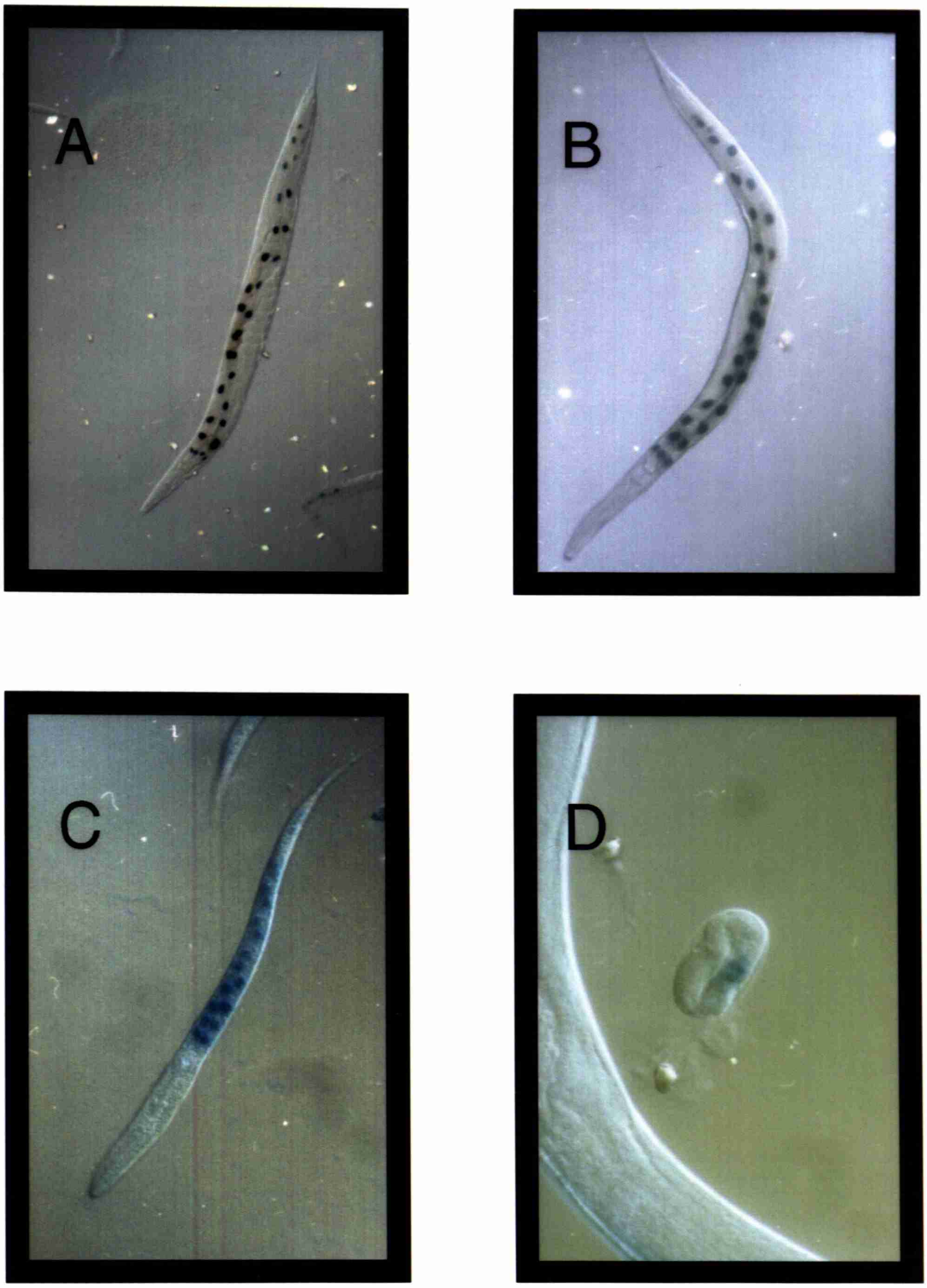

A
B
C
D

Figure 5.25

Expression of the *cpr-5*::*lacZ* transgene fusion in transgenic *C.elegans*

Strains of *C.elegans* transgenic for the *cpr-5*::*lacZ* transgene fusion were fixed and stained for ß-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C.elegans* development.

A. Adult

B. L_3

C. L_2

D. Embryo (threefold stage)

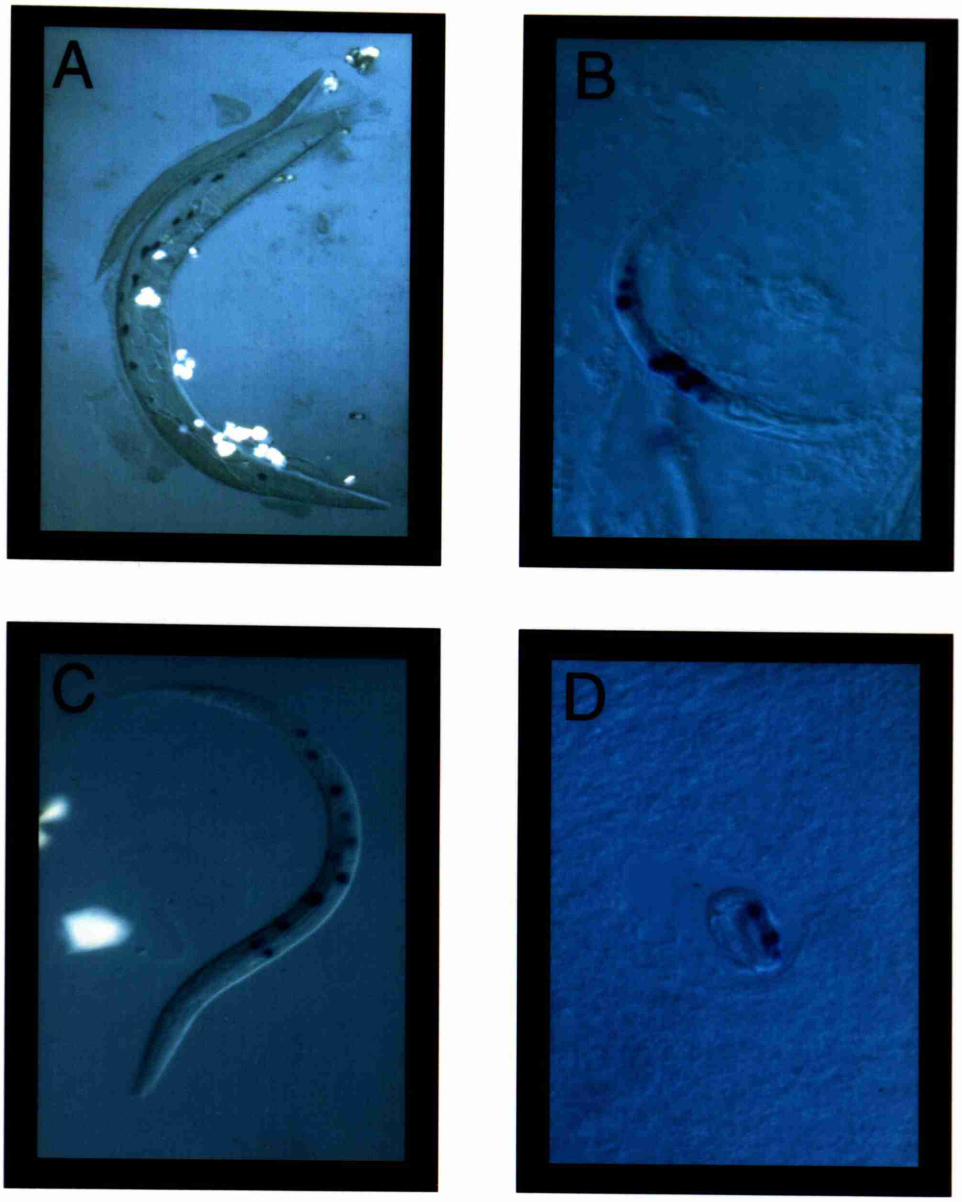

A
B
C
D

Figure 5.26

Expression of the *cpr-6*::*lacZ* transgene fusion in transgenic *C.elegans*

Strains of *C.elegans* transgenic for the *cpr-6*::*lacZ* transgene fusion were fixed and stained for ß-galactosidase activity with X-gal. The figure shows representative examples of the staining patterns obtained for different stages of *C.elegans* development.

A. Adult

B. Two L_4 larvae

C. L_3

D. L_2

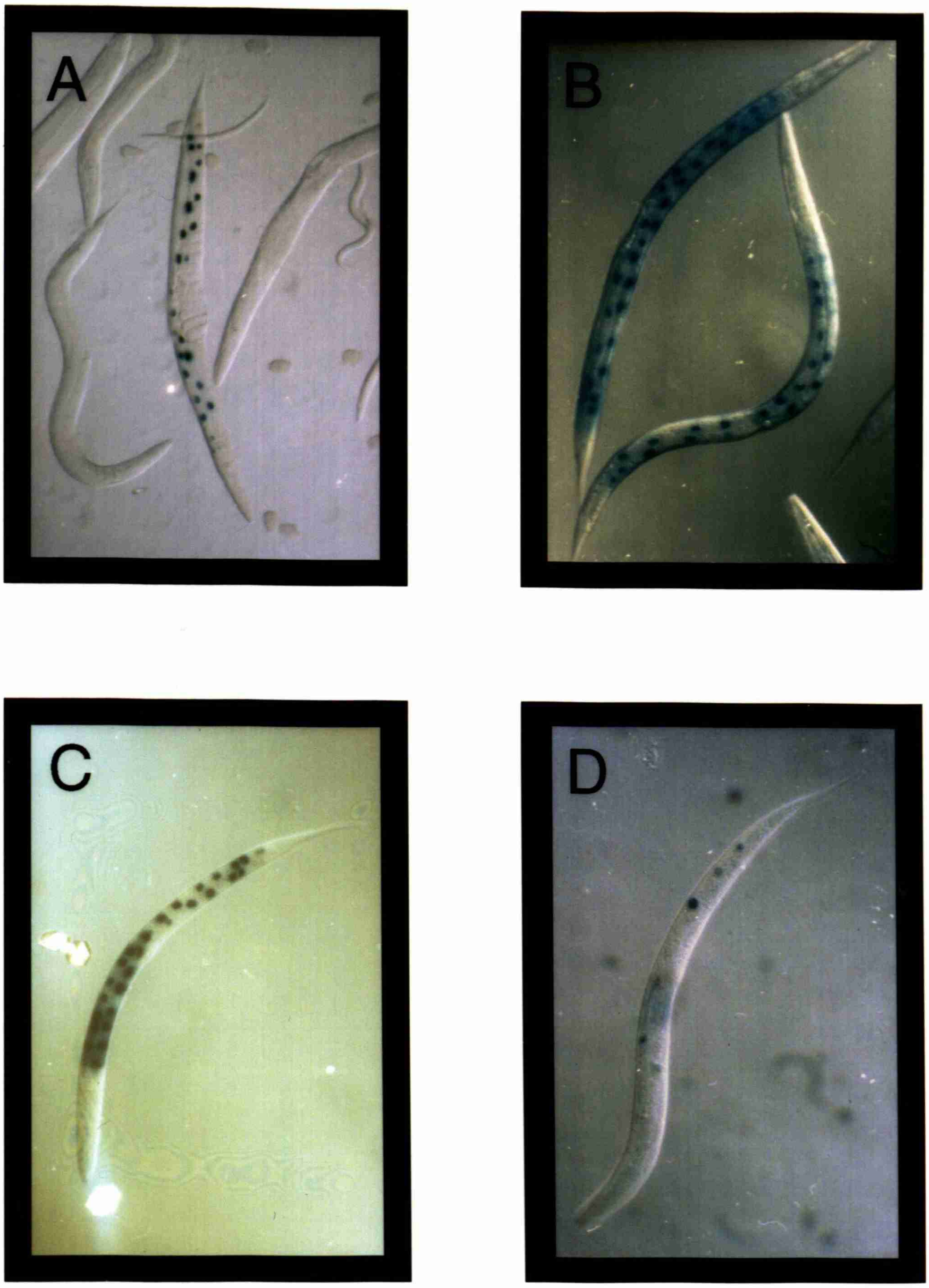
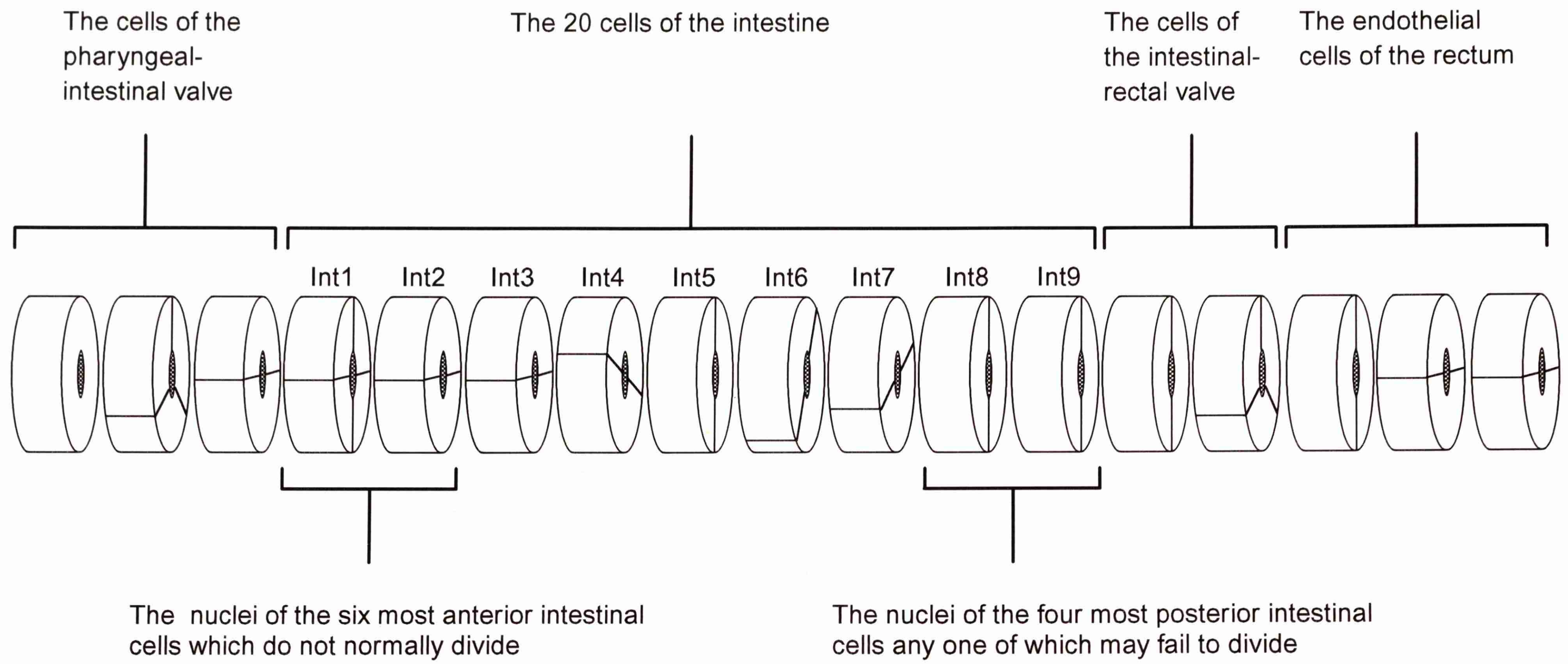

A
B
C
D

Figure 5.27

Schematic diagram of part of the alimentary tract of the early L_1.

The figure shows the 20 intestinal cells of the gut and their arrangement into nine groups (int1-int9). The ring of four anterior most intestinal cells (int1) are attached to the pharyngeal-intestinal valve. The remaining 16 intestinal cells are arranged as eight pairs (int2-int9). During L1 lethargus most of the intestinal cells undergo nuclear divisions to generate binucleated cells. However, the nuclei of the six anterior most intestinal cells (int1 and int2) do not normally divide and any of the nuclei of the four posterior most intestinal cells (int8 and int9) may also fail to divide. Thus, the intestinal cells of L2 larvae (and all subsequent developmental stages) normally possess 30-34 nuclei. The nuclei of all 20 intestinal cells undergo a single endoreduplication event at each moult resulting in C numbers of 32 in the adult.

The figure was adapted from Sulston *et al* (1983).


The cells of the pharyngeal-intestinal valve
The 20 cells of the intestine
The cells of the intestinal-rectal valve
The endothelial cells of the rectum
Int1
Int2
Int3
Int4
Int5
Int6
Int7
Int8
Int9
The nuclei of the six most anterior intestinal cells which do not normally divide
The nuclei of the four most posterior intestinal cells any one of which may fail to divide

Chapter 6

Chapter 6

Concluding Discussion

6.1. Summary of results

The work described in this thesis indicates that *C.elegans* possesses a cathepsin B-like multigene family comprising at least five members, the four cathepsin B-like genes *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* reported here and the previously isolated gene *cpr-1* (Ray and McKerrow, 1992). DNA sequencing of the genomic, cDNA and 5′ RACE clones revealed that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* all possess diverged gene architectures, suggesting that they have arisen from ancient gene duplication events.

Some genes in *C.elegans* appear to be transcribed polycistronically in clusters, similar to the situation observed for bacterial operons (Spieth *et al.,* 1993). Mature, monocistronic messages are generated from these polycistronic pre-mRNAs by trans-splicing events and the spliced leader SL2 is specific for trans-splicing to downstream genes in these operons (Spieth *et al.,* 1993). A survey of DNA sequence data generated by the *C.elegans* genome project identified 31 potential operons (Zorio *et al.,* 1994). The authors analysed 2 Mb of DNA sequence for the presence of trans-splice sites and for gene clusters. This analysis suggested that over 70% of the 266 predicted genes examined were trans-spliced and that 26% of these genes were in operons. The authors analysed seven of these potential operons for trans-splicing and found that all fitted the expectations for polycistronic transcription. RT-PCR was used to determine whether the predicted genes were transcribed and the mRNAs transpliced. These experiments revealed that the first gene of each of the putative operons tested was either trans-spliced to SL1 or not trans-spliced at all. In contrast, downstream genes were trans-spliced exclusively to SL2 or to a mixture of SL2 and SL1.

The 5′ RACE approach was used to study the transcripts of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. The 5′ RACE experiments indicated that none of these genes are trans-spliced. Thus, none of these four genes represent downstream genes in an operon and therefore, none are transcribed from the same operon. The 5′ RACE experiments indicated that transcription initiation occurs within a small region upstream of and proximal to the putative ATG translation initiation codon of each gene, resulting in the mature transcripts of these genes possessing short 5′ UTRs. These results also suggest that each of the four genes is under the control of a single promoter region proximal to the 5′ end of the respective gene.

Northern analysis detected a single size species of mature transcript for each gene. Each species of mature transcript was of the expected size for its respective gene. For each of the four genes, no evidence for alternative splicing was obtained within the limits of resolution and detection of Northern analysis. The results obtained from the 5′ RACE experiments and from Northern analysis are in agreement with those obtained for *cpr-1* (Ray and McKerrow, 1992). Northern analysis identified a single size species of transcript for *cpr-1* and primer extension analysis indicated that the mature *cpr-1* transcript possesses a short 5′ UTR which is not trans-spliced.

The predicted amino acid sequences encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* display all the characteristics expected of a cathepsin B-like enzyme. In addition, phylogenetic analysis of the predicted amino acid sequences encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* clusters these sequences with known cathepsin B enzymes and away from similar but distinct cysteine proteases such as cathepsin L and cathepsin H. Together, these data support the assignment of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* as cathepsin B-like genes.

Though the putative enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* possess features characteristic of the cathepsin B enzyme, most of these predicted proteins are highly diverged from one another. This is demonstrated by phylogenetic analysis which organises these five predicted proteins into three groups and indicates that these three groups are almost as diverged from one another as each is from the vertebrate cathepsin B enzymes. Indeed, of the five putative enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*, only those encoded by *cpr-4* and *cpr-5* show substantially more homology to one another than to human cathepsin B. Furthermore, comparison of the predicted amino acid sequences encoded by these five genes to human cathepsin B, reveals a number of variations at amino acid residues shown to affect the specificity and activity of recombinant rat cathepsin B (Hasnian *et al.*, 1992). These results suggest that most of the putative enzymes encoded by *cpr-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are highly diverged from one another and may have enzyme activities and substrate specificities both distinct from human cathepsin B and distinct from each other.

The work also strongly suggests that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are all exclusively expressed in the intestinal cells of *C.elegans*. This is in agreement with *in-situ* hybridisation experiments (Ray and McKerrow, 1992) and *lacZ* transgene fusion experiments (C.Britton, pers. comm.) conducted with *cpr-1* which suggest that expression of this gene is also confined to the intestinal cells of *C.elegans*. Though *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are all expressed in the intestine, they have different patterns of temporal expression (Larminie and Johnstone, 1996). Thus, the results suggest that expression of *cpr-3*, *cpr-4* and *cpr-5* is induced at the late stages of embryogenesis but that these three genes show different temporal expression patterns from one another in the subsequent larval and adult stages of development. *cpr-3* relative transcript abundance remains fairly constant post hatch, suggesting that this gene may be constitutively expressed. *cpr-4* relative transcript abundance is elevated during the larval stages with peaks in L_1 and L_4, but decline significantly in the adult by comparison to *cpr-3*, suggesting that *cpr-4* expression is upregulated during the larval stages of *C.elegans* development. *cpr-5* relative transcript abundance increases dramatically post hatch, reaches a peak at the L_2 stage, remains high in the two subsequent larval stages and declines slightly in the adult. In contrast to *cpr-3*, *cpr-4* and *cpr-5*, the relative transcript abundance of *cpr-6* does not increase dramatically until much later in development, around the L_3-L_4 stages, and continues to rise through the early adult stages.

The results suggest that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are all expressed exclusively in the intestinal cells of *C.elegans* and possess distinct, but overlapping temporal patterns of expression. The predicted amino acid sequences of the four genes indicate that they encode cathepsin B-like enzymes which may have different activities and substrate specificities. Together, these data suggest a requirement for different cathepsin B-like enzymes, or combinations of such enzymes, to process or degrade different substrates, or combinations of substrates, in the intestine during different stages of *C.elegans* development. Since the gut of *C.elegans* performs an important role in digestion and is also thought to be the principle storage organ of the worm, the four cathepsin B-like enzymes may be involved in digestion and/or the degradation/processing of stored proteins. Thus, some or all of these enzymes may be directly involved in digestion by degrading proteins in the gut lumen. The different activities and substrate specificities proposed for these enzymes may be important in this respect since the presence of more than one cathepsin B-like enzyme in the gut lumen would increase the spectrum of protein substrates which can be degraded. Alternatively, some or all of the enzymes may perform indirect, intracellular roles in digestion, by processing other enzymes that are required for digestion in the gut lumen. *cpr-3*, *cpr-4* and *cpr-5* encode the most likely candidates for digestive enzymes since these genes are all expressed throughout the larval and adult stages of *C.elegans* development (the stages when this nematode species feeds).

Since the intestine of *C.elegans* is also thought to represent the principle storage organ of the worm, some or all of these enzymes may be involved in the degradation or processing of stored proteins. In this respect, the distinct but overlapping temporal expression patterns of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* may be significant since different combinations of cathepsin B-like enzymes may be required for the processing/degradation of different stored proteins at different stages during the life cycle of *C.elegans*. The putative cathepsin B-like enzyme encoded by *cpr-6* may be a more likely candidate for such a role because this gene is not expressed at significant levels in all the feeding stages of *C.elegans* development and is therefore less likely to have an essential role in digestion. However, since roles in digestion and roles in protein processing are not mutually exclusive, some or all of these enzymes may perform dual functions. For example, one cathepsin B-like enzyme may process enzymes required for digestion but also process other stored proteins. Again, the different temporal patterns of expression observed for the four genes may be significant since different combinations of enzymes with different activities and substrate specificities may be able to effect different but related functions during *C.elegans* development. Thus one combination of enzymes may effect a digestive role while another combination may effect a role in the degradation/processing of stored proteins.

The identification of five diverged cathepsin B-like genes that are all expressed in the gut and that have distinct but overlapping temporal patterns of expression suggests that several different cathepsin B-like enzymes may be active within the same cell at any one time during *C.elegans* post-embryonic development. The protein substrates hydrolysed by these enzymes and the nature of the hydrolysis (i.e. complete degradation or limited proteolysis of a substrate) may be dependent on the combination of active enzymes present. Since at least five cathepsin B-like genes may be expressed in the same intestinal cells, a large number of potential enzyme combinations are possible and many of these combinations may be functionally significant. However the potential number of enzyme combinations may be significantly reduced when one considers where the final active enzymes are located. Thus, cathepsin B-like enzymes which are secreted into the gut lumen are unlikely to have overlapping functions with those that reside within the intestinal cells. It is also possible that not all intracellular cathepsin B-like enzymes will have overlapping functions since they may not be targeted to the same organelles. This is perhaps less likely because cathepsin B-like enzymes normally have acidic pH optima and are therefore associated with the lysosomal/endosomal compartment (or specialised organelles which are associated with this compartment, such as secretory vesicles). However, since we have not biochemically characterised these enzymes nor studied their subcellular distribution, such a possibility cannot be discounted.

The presence of a cathepsin B-like multigene family in *C.elegans* whose members exhibit spatially restricted patterns of expression is similar to the situation observed in parasitic nematode and trematode species, where cathepsin B-like multigene families have been identified (Chapter 1, Section 1.9) and evidence of tissue-restricted patterns of expression obtained (Chapter 1, Section 1.8). By comparison, vertebrates appear to possess only a single cathepsin B gene that is expressed in many, if not all, tissues of the body. Furthermore, unlike the cathepsin B-like genes of *C.elegans*, the vertebrate cathepsin B gene appears to be under the control of a complex regulatory mechanism, with evidence of differential expression from more than one promoter and of alternative splicing of the resulting transcripts. Together, these observations suggest that the cathepsin B-like enzymes of both free living and parasitic nematode and trematode species may be required for different biological functions than the cathepsin B enzyme of vertebrates.

The identification of cathepsin B-like multigene families in both a free-living nematode species and in parasitic nematode and trematode species does not preclude such enzymes performing parasite-specific roles since the biological roles of such enzymes in free-living nematode species may have been adapted for a parasitic lifestyle. However, it does indicate that the amplification of cathepsin B-like multigene families is not unique to parasitic nematode and trematode species and therefore indicates that these enzymes may be required for general nematode and trematode biology, and not just for parasite-specific processes.

6.2. Future work

6.2.1. Determining the distribution of the cathepsin B-like enzymes

In order to determine the distribution of cathepsin B-like enzymes in *C.elegans*, it would be necessary to generate antisera specific for each of the enzymes. Specific antisera may be obtained by one of two routes. The first route would be to use synthetic peptides to immunise rabbits. There are several advantages to using such an approach to obtain antisera specific for each cathepsin B-like enzyme. The extensive knowledge of the human cathepsin B enzyme tertiary structure could be used to predict the tertiary structure of each cathepsin B-like enzyme. This would allow more accurate prediction of synthetic peptides likely to be immunogenic for each of the cathepsin B-like enzymes. Such an approach would also allow the rational selection of synthetic peptides from regions which are not conserved between each of the enzymes and so increase the probability of obtaining specific antisera for each cathepsin B-like enzyme. The second approach would be to express the cDNA for each gene in *E.coli* and use the expressed proteins to immunise rabbits. Such an approach has the disadvantage that it is likely to produce antisera for each enzyme which will cross react with other members of the cathepsin B-like multigene family. This may be overcome by expressing only portions of each cDNA which encode regions of variability between the four enzymes.

Were specific antisera to be obtained, there are a number of studies which could be performed. First, the antisera could be used for whole worm antibody staining. Such experiments would allow confirmation of the data obtained with the *lacZ* fusion transgenes and allow analysis of the distribution of the protein products of the four genes. In the latter case, this could determine which, if any, of the four genes may encode cathepsin B-like enzymes with a direct or indirect role in digestion by demonstrating their presence in the lumen of the intestine. Such studies would not allow cathepsin B-like enzymes with a direct role in digestion to be distinguished from those with indirect roles since both types of role are likely to result in the presence of such enzymes in the lumen. Cathepsin B-like enzymes with a direct role in digestion will be localised in the intestinal lumen for obvious reasons. However, cathepsin B-like enzymes that perform indirect roles, by processing proteins required for digestion, may perform such roles in secretory granules and may therefore be released into the lumen with the processed proteins.

Since the results suggest that *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* are all expressed in the intestine of *C.elegans*, it would be informative to use specific antisera for immunogold electron microscopy. This approach would enable the subcellular localisation of each enzyme to be studied in detail. Furthermore, by using different sizes of gold particles, it would be possible to determine which, if any, of these enzymes are co-localised to the same organelles. The results from such studies could thereby identify cathepsin B-like enzymes which may perform overlapping functions within the same subcellular compartment.

6.2.2. Biochemical characterisation of the cathepsin B-like enzymes

Specific antisera for each of the four cathepsin B-like enzymes may also allow purification of the enzymes by antibody affinity chromatography. The free living existence of *C.elegans* means there is no real limit to the amount of material which can be obtained and therefore such an approach may allow the purification of sufficient quantities of each enzyme for extensive biochemical characterisation. Biochemical characterisation of the proteins encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* will form an important part of any future work. The proteins encoded by these genes have been termed cathepsin B-like by virtue of their homology to human cathepsin B and because they possess many of the amino acid sequence characteristics expected of cathepsin B-like enzymes. However, biochemical characterisation is essential in order to prove that these encoded proteins have cathepsin B-like activities. First, the activities and substrate preferences of the four enzymes would be studied using both protein substrates and fluorogenic synthetic peptide substrates. Protein substrates would be used to determine whether these enzymes do genuinely have different activities and substrate specificities, as suggested by their predicted amino acid sequences. Most notably, this approach would determine which of the enzymes possess the peptidyldipeptidase activity associated with vertebrate cathepsin B. Banks of fluorogenic synthetic peptide substrates possessing different amino acid combinations would be used to analyse the amino acid substrate preferences of the four enzymes. In particular, the ability of these enzymes to cleave synthetic substrates containing two basic amino acid residues would be tested, since cleavage of such substrates represents another characteristic feature of vertebrate cathepsin B. In addition, the screening of such banks might reveal synthetic substrates which are specific for each enzyme. These specific synthetic substrates would allow studies similar to those performed with parasitic nematode species. Thus the E/S products from liquid cultures of *C.elegans* and the proteins from whole worm homogenates could be analysed using the specific synthetic substrates. This might identify which enzymes are active outwith *C.elegans* and therefore might identify which, if any, of these cathepsin B-like enzymes are likely to have a direct role in digestion. It should be noted that such studies would also require the isolation of the enzyme encoded by the *cpr-1* gene (Ray and McKerrow, 1992) so that it could be tested to ensure that it did not possess any significant activity against the chosen synthetic substrates. Even after such studies, the results would have to be interpreted with some caution because one could not exclude the possibility of the presence of additional uncharacterised cathepsin B-like enzymes in *C.elegans* with similar synthetic substrate preferences to the enzymes being analysed.

The second area of biochemical characterisation would be to analyse the enzyme kinetics of each of the proteins encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*. Fluorogenic substrates would be used to determine the pH at which each enzyme shows optimal activity and the pH range within which each enzyme is active. Cysteine protease inhibitors would be used to demonstrate that these enzymes are indeed cysteine proteases. Small synthetic peptide inhibitors would also be used to study the rates of inactivation of each enzyme. The specificity of these inhibitors is due to the peptide portion of the inhibitor which is bound by the subsites of the active site cleft. Thus, by screening banks of inhibitors which possess different combinations of amino acid residues, it would be possible to analyse the subsite preferences of each enzyme. Such a study would complement the results obtained from screening banks of fluorogenic synthetic peptide substrates.

Expression of recombinant rat cathepsin B in yeast results in the generation of active enzyme (Lee *et al.,* 1990) and this enzyme has been shown to be essentially identical to rat liver cathepsin B by detailed biochemical analysis (Hasnain *et al.,* 1992). Accordingly, it may also be possible to obtain active recombinant cathepsin B-like enzymes by expressing the cDNA of each of the four genes in yeast. This system has the advantage that it does not require specific antisera and thus would allow biochemical characterisation of the cathepsin B-like enzymes encoded by *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* regardless of whether or not specific antisera against these enzymes could be obtained. However, since no biochemical comparisons between the recombinant enzymes and the endogenous enzymes could be made, the results from such studies would have to be treated with a certain degree of caution.

6.2.3. The effect of starving on expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The gut-specific expression of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* suggests that these genes may be involved in digestion. Thus expression of at least some of these genes may only be induced in the presence of a food source. This aspect of gene regulation could be studied by repeating the s-q rtPCR experiments using RNA isolated from embryos and L_1 larvae cultured in the presence and absence of an *E.coli* food source. The identification of cathepsin B-like genes which are only expressed in the presence of a food source would provide strong evidence for a role in digestion for such genes. By the same token, it would also be useful to use s-q rtPCR to determine whether or not *cpr-3, cpr-4*, *cpr-5* and *cpr-6* are expressed in the non-feeding dauer larvae. Such studies would also demonstrate the presence of complex regulatory mechanisms which ensure that expression of such genes only occurs in the appropriate conditions.

6.2.4. Obtaining *C.elegans* mutant strains for *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*

The final area of future research highlights the experimental tractability of *C.elegans*. Strains of *C.elegans* with mutations in *cpr-3*, *cpr-4*, *cpr-5* or *cpr-6* may be obtained by one of two routes. First, these genes may already be defined by genetic loci with known genetic map positions, determined by classical genetic analysis.

Accordingly, the alignment of the physical and genetic maps of *C.elegans* may allow the isolation of mutant strains for some or all of these genes by plasmid rescue with the cloned wild-type genes (Chapter 3, Section 3.3). If such an approach failed, mutant strains could be obtained using a reverse genetic approach. This would be achieved using the site-selected transposon mutagenesis technique outlined in Chapter 1, Section 1.3. The mutants for each of the four genes obtained by these approaches may possess phenotypes which will help to determine the biological roles for each of the four genes. Alternatively, such mutants may not possess a visible mutant phenotype. This may be a result of functional redundancy within the cathepsin B-like multigene family. Such redundancy could be analysed by constructing crosses to generate strains with mutations in two or more of the cathepsin B-like genes. Such studies could not only identify genes with overlapping functions but also gene combinations essential for nematode survival. This information may be particularly useful for the effective design of chemotherapeutic agents against parasitic nematode species.

References

References

Ahlberg, J., Berkenstam, A., Henell, F., and Glaumann, H. (1985) Degradation of short and long lived proteins in isolated rat liver lysosomes. *Journal of Biological Chemistry* **260**, 5847-5854.

Aird, W.C., Parvin, J.D., Sharp, P.A., and Rosenberg, R.D. (1994) The interaction of GATA-binding proteins and basal transcription factors with GATA box-containing core promoters: a model of tissue specific gene expression. *Journal of Biological Chemistry* **269**, 883-889.

Aroian, R.V., Koga, M., Mendel, J.E., Ohshima, Y., and Sternberg, P.W. (1990) The *let-23* gene necessary for *Caenorhabditis elegans* vulvar induction encodes a tyrosine kinase of the EGF receptor subfamily. *Nature* **348**, 693-699.

Aronson, N.N. and Barrett, A.J. (1978) The specificity of cathepsin B. *Biochemical Journal* **171**, 759-765.

Baici, A. and Knopfel, M. (1986) Cysteine proteinases produced by cultured rabbit V2 carcinoma cells and rabbit skin fibroblasts. *International Journal of Cancer* **38**, 753-761.

Barnes, R.D. (1980) *Invertebrate Zoology*, Fourth Ed., 263-315, Philadelphia, Holt-Saunders.

Barrett, A.J. (1977) Cathepsin B and Other Thiol Proteinases. In: *Proteinases in Mammalian Cells and Tissues*, 181-208. Edited by Barrett, A.J. Amsterdam, North-Holland Publishing Co.

Barrett, A.J., Kembhavi, A.A., Brown, M.A., Kirschke, H., Knight, C.G., Tamai, M., and Hanada, K. (1982) L-trans-epoxysuccinyl-leucylamido(4-guanidino)butane (E-64) and its analogs as inhibitors of cysteine proteinases including cathepsins B, H and L. *Biochemical Journal* **201**, 189-198.

Barrett, A.J. and Kirschke, H. (1981) Cathepsin B, cathepsin H and cathepsin L. *Methods in Enzymology* **80**, 535-561.

Beitel, G., Clark, S., and Horvitz, H.R. (1990) *Caenorhabditis elegans ras* gene *let-60* acts as a switch in the pathway of vulval differentiation. *Nature* **348**, 503-509.

Berquin, I, Cao, L., Fong, D., and Sloane, B.F. (1995) Identification of two new exons and multiple transcription start points in the 5'-untranslated region of the human cathepsin-B-encoding gene. *Gene* **159**, 143-149.

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

Blaxter, M.L. (1993) Nemoglobins: Divergent nematode globins. *Parasitology Today* **9**, 353-360.

Blumenthal, T., Squire, M., Kirtland, S., Cane, J., Donegan, M., Spieth, J., and Sharrock, W. (1984) Cloning of a yolk protein gene family from *Caenorhabditis elegans. Journal of Molecular Biology* **174**, 1-18.

Bond, J.S. and Barrett, A.J. (1980) Degradation of fructose-1,6-bisphosphate aldolase by cathepsin B: a further example of peptidyldipeptidase activity of this enzyme. *Biochemical Journal* **189**, 17-25.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans. Genetics* **77**, 71-94.

Burleigh, M.C., Barrett, A.J., and Lazarus, G.S. (1974) Cathepsin BI, A lysosomal enzyme that degrades native collagen. *Biochemical Journal* **137**, 387-398.

Cangiano, G. and La Volpe, A. (1993) Repetitive DNA sequences located in the terminal portion of the *Caenorhabditis elegans* chromosomes. *Nucleic Acids Research* **21**, 1133-1139.

Cao, L., Taggart, R.T., Berquin, I.M., Moin, K., Fong, D., and Sloane, B.F. (1994) Human gastric adenocarcinoma cathepsin B: isolation and sequencing of full-length cDNAs and polymorphisms of the gene. *Gene* **139**, 163-169.

Casey, J.L., Hentze, M.W., Koeller, D.M., Caughman, S.W., Rouault, T.A., Klausner, R.D., and Harford, J.B. (1988) Iron-responsive elements: regulatory RNA sequences that control messenger-RNA levels and translation. *Science* **240**, 924-928.

Cazzulo, J.J., Martínez, J., Parodi, A.J.A., Wernstedt, C., and Hellman, U. (1992) On the post-translational modifications at the C-terminal domain of the major cysteine proteinase (cruzipain) from *Trypanosoma cruzi*. *FEMS Microbiology Letters* **100**, 411-416.

Chalfie, M. and White, J. (1988) The Nervous System. In: *The Nematode Caenorhabditis elegans*, 337-391. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Chan, S.J., Sansegundo, B., McCormick, M.B., and Steiner, D.F. (1986) Nucleotide and predicted amino acid sequences of cloned human and mouse preprocathepsin B cDNAs. *Proceedings of the National Academy of Sciences (USA)* **83**, 7721-7725.

Chance, R.E. (1972) Amino acid sequences of proinsulin and intermediates. *Diabetes* **21 (suppl. 2)**, 461-467.

Clark, L. and Carbon, J. (1976) A colony bank containing synthetic Col E1 hybrid plasmids representative of the entire *E.coli* genome. *Cell* **9**, 91-99.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences (USA)* **83**, 7821-7825.

Coulson, A., Waterston, R., Kiff, J., Sulston, J., and Kohara, Y. (1988) Genome linking with yeast artificial chromosomes. *Nature* **335**, 184-186.

Coulson, A., Kozono, Y., Lutterbach, B., Shownkeen, R., Sulston, J., and Waterston, R. (1991) YACs and the *C.elegans* genome. *BioEssays* **13**, 413-417.

Cox, G.N., Pratt, D., Hageman, R., and Boisvenue, R.J. (1990) Molecular cloning and primary sequence of a cysteine protease expressed by *Haemonchus contortus* adult worms. *Molecular and Biochemical Parasitology* **41**, 25-34.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978) A Model of Evolutionary Change in Proteins. In: *Atlas of Protein Sequence and Structure*, 345-352. Edited by Dayhoff, M. Washington D.C. National Biomedical Research Foundation.

Dean, R.T. (1975) Lysosomal enzymes as agents of turnover of soluble proteins. *European Journal of Biochemistry* **58**, 9-14.

Delaisse, J.-M., Eeckhout, Y., and Vaes, G. (1980) inhibition of bone resorption in culture by inhibitors of thiol proteases. *Biochemical Journal* **192**, 365-368.

Delaisse, J.-M., Eeckhout, Y., and Vaes, G. (1984) *In vivo* and *in vitro* evidence for the involvement of cysteine proteinases in bone resorption. *Biochemical and Biophysical Research Communications* **125**, 441-447.

Delaisse, J.-M., Ledent, P., and Vaes, G. (1991) Collagenolytic cysteine proteinases of bone tissue. *Biochemical Journal* **279**, 167-174.

Docherty, K., Carroll, R., and Steiner, D.F. (1983) Identification of a 31,500 molecular weight islet cell protease as cathepsin B. *Proceedings of the National Academy of Sciences (USA)* **80**, 3245-3249.

Docherty, K. and Steiner, D.F. (1982) Post-translational proteolysis in polypeptide hormone biosynthesis. *Annual Review Of Physiology* **44**, 625-638.

Doolittle, W.F. and Stoltzfus, A. (1993) Molecular evolution: genes-in-pieces revisited. *Nature* **361**, 403

Drenth, J., Jansonius, J.N., Koeboek, R., Swen, H.M., and Wothers, B.G. (1968) Structure of papain. *Nature* **218**, 929-932.

Drenth, J., Jansonius, J.N., Koeboek, R., and Wolthers, B.G. (1971) Papain, X-Ray Structure. In: *The Enzymes*, Third Ed., 485-499. Edited by Boyer, P.D. London, Academic Press.

Dresden, M.H., Rege, A.A., and Murrell, K.D. (1985) *Strongyloides ransomi:* proteolytic enzymes from larvae. *Experimental Parasitology* **59**, 257-263.

Dunn, A.D., Crutchfield, H.E., and Dunn, J.T. (1991a) Thyroglobulin processing by thyroidal proteases. *Journal of Biological Chemistry* **266**, 20198-20204.

Dunn, A.D., Crutchfield, M.E., and Dunn, J.T. (1991b) Proteolytic processing of thyroglobulin by extracts of thyroid lysosome. *Endocrinology* **128**, 3973-3080.

Dunn, B.M. (1989) Determination of protease mechanism. In: *Proteolytic Enzymes, A Practical Approach*, 57-81. Edited by Beynon, R.J. and Bond, J.S. Oxford, IRL Press.

Edwards, M.K. and Wood, W.B. (1983) Location of specific messenger RNAs in *Caenorhabditis elegans* by cytological hybridisation. *Developmental Biology* **97**, 375-390.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* **37**, 36-48.

Egan, C.R., Chung, M.A., Allen, F.L., Heschl, M.F.P., van Buskirk, C.L., and McGhee, J.D. (1995) A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans ges-1* gene centers on two GATA sequences. *Molecular and Cellular Biology*

Emmons, S.W. (1988) The Genome. In: *The Nematode Caenorhabditis elegans*, 47-79. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Evans, T., Reitman, M., and Felsenfeld, G. (1988) An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proceedings of the National Academy of Sciences (USA)* **85**, 5976-5980.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package). (Abstract)

Ferrara, M., Wojcik, F., Rhaissi, H., Mordier, S., Roux, M.P., and Bechet, D. (1990) Gene structure of mouse cathepsin B. *FEBS Letters* **273**, 195-199.

Files, J.G. and Hirsh, D. (1981) Ribosomal DNA of *Caenorhabditis elegans. Journal of Molecular Biology* **149**, 223-240.

Fire, A. (1986) Integrative transformation of *Caenorhabditis elegans. EMBO Journal* **5**, 2673-2680.

Fire, A., White Harrison, S., and Dixon, D. (1990) A modular set of *lacZ* fusion vectors for studying gene expression in *Caenorhabditis elegans. Gene* **93**, 189-198.

Fire, A., Albertson, D., Harrison, S.W., and Moerman, D.G. (1991) Production of antisense RNA leads to effective and specific inhibition of gene expression in *C.elegans* muscle. *Development* **113**, 503-514.

Fong, D., Calhoun, D.H., Hsieh, W.-T., Lee, B., and Wells, R.D. (1986) Isolation of a cDNA clone for the human lysosomal proteinase cathepsin B. *Proceedings of the National Academy of Sciences (USA)* **83**, 2909-2913.

Fox, T., de Miguel, E., Mort, J.S., and Storer, A.C. (1992) Potent slow-binding inhibition of cathepsin B by its propeptide. *Biochemistry* **31**, 12571-12576.

Glazer, A.N. and Smith, E.L. (1971) Papain and Other Plant Sulfhydryl Proteolytic Enzymes. In: *The Enzymes*, Third Ed., 501-546. Edited by Boyer, P.D. London, Academic Press.

Gong, Q., Chan, S.J., Bajkowski, A.S., Steiner, D.F., and Frankfater, A. (1993) Characterization of the cathepsin B gene and multiple mRNAs in human tissues: evidence for alternative splicing of cathepsin B pre-mRNA. *DNA and Cell Biology* **12**, 299-209.

Gotz, B. and Klinkert, M.-Q. (1993) Expression and partial characterisation of a cathepsin B-like enzyme (Sm31) and a proposed haemoglobinase (Sm32) from *S.mansoni*. *Biochemical Journal* **290**, 801-806.

Greenbaum, L.M. and Fruton, J.S. (1957) Purification and properties of beef spleen cathepsin B. *Journal of Biological Chemistry* **226**, 173-180.

Guagliardi, L.E., Koppelman, B., Blum, J.S., Marks, M.S., Cresswell, P., and Brodsky, F.M. (1990) Co-localization of molecules involved in antigen processing and presentation in an early endocytic compartment. *Nature* **343**, 133-139.

Guo, S. and Kemphues, K.J. (1995) *par-1*, a gene required for establishing polarity in *C.elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell* **81**, 611-620.

Hamelin, M., Scott, I.M., Way, J.C., and Culotti, J. (1992) The *mec-7* b-tubulin gene of *Caenorhabditis elegans* is expressed primarily in the touch receptor neurons. *EMBO Journal* **11**, 2885-2893.

Han, M., Golden, J., Han, Y., and Sternberg, P.W. (1993) *C.elegans lin-45 raf* gene participates in *let-60 ras*-stimulated vulval differentiation. *Nature* **363**, 133-139.

Han, M. and Sternberg, P.W. (1990) *let-60*, a gene that specifies cell fates during *Caenorhabditis elegans* vulval induction, encodes a *ras* protein. *Cell* **63**, 921-931.

Hara, K., Kominami, E., and Katanuma, N. (1988) Effect of proteinase inhibitors on intracellular processing of cathepsins B,H and L in rat macrophages. *FEBS Letters* **231**, 229-231.

Harrop, S.A., Sawangjaroen, N., Prociv, P., and Brindley, P.J. (1995) Characterisation and localisation of cathepsin B proteinases expressed by adult *Ancyclostoma caninum* hookworms. *Molecular and Biochemical Parasitology* **71**, 163-171.

Hasnain, S., Hirama, T., Tam, A., and Mort, J.S. (1992) Characterization of recombinant rat cathepsin B and nonglycosylated mutants expressed in yeast: new insights into the pH-dependence of cathepsin B-catalyzed hydrolyzes. *Journal of Biological Chemistry* **267**, 4713-4721.

Hasnian, S., Huber, C.P., Muir, A., Rowan, A.D., and Mort, J.S. (1992) Investigation of structure function relationships in cathepsin B. *Biological Chemistry Hoppe-Seyler* **373**, 413-418.

Hawkins, M.G. and McGhee, J.D. (1995) *elt-2*, a second GATA factor from the nematode *Caenorhabditis elegans*. *Journal of Biological Chemistry* **270**, 14666-14671.

Hawthorne, J., Halton, D.W., and Walker, B. (1993) Identification and characterization of the cysteine and serine proteases of the trematode, *Haplometra cylindracea* and determination of their haemoglobinase activity. *Parasitology* **108**, 595-601.

Hedgecock, E.M. and White, J.G. (1985) Polyploid tissues in the nematode *Caenorhabditis elegans*. *Developmental Biology* **107**, 128-133.

Hengartner, M.O. and Horvitz, H.R. (1994) *C.elegans* cell survival gene *ced-9* encodes a functional homolog of the mammalian protooncogene *bcl-2*. *Cell* **76**, 665-676.

Heussler, V.T. and Dobbelaere, D.A.E. (1994) Cloning of a protease gene family of *Fasciola hepatica* by the polymerase chain reaction. *Molecular and Biochemical Parasitology* **64**, 11-23.

Higgins, D.G., Bleasby, A.J., and Fuchs, R. (1992) Clustal V: Improved software for multiple sequence alignment. *Computer Applications in the Biosciences* **8**, 189-191.

Hill, P.A., Buttle, D.J., Jones, S.J., Boyde, A., Murata, M., Reynolds, J.J., and Meikle, M.C. (1994) Inhibition of bone resorption by selective inactivators of cysteine proteinases. *Journal Of Cellular Biochemistry* **56**, 118-130.

Hill, R.J. and Sternberg, P.W. (1992) The gene *lin-3* encodes an inductive signal for vulval development in *C.elegans*. *Nature* **358**, 470-476.

Hodgkin, J., Horvitz, H.R., and Brenner, S. (1979) Nondisjunction mutants of the nematode *Caenorhabditis elegans*. *Genetics* **91**, 67-94.

Hopgood, M.F., Clark, M.G., and Ballard, F.J. (1977) Inhibition of protein degradation in isolated rat hepatocytes. *Biochemical Journal* **164**, 399-407.

Howie, A.J., Burnett, D., and Crocker, J. (1985) The distribution of cathepsin B in human tissues. *Journal of Pathology* **145**, 307-314.

Karrer, K.M., Peiffer, S.L., and DiTomas, M.E. (1993) Two distinct subfamilies within the family of cysteine protease genes. *Proceedings of the National Academy of Sciences (USA)* **90**, 3063-3067.

Keppler, D., Fondaneche, M.C., Dalet-Fumaron, V., Pagano, M., and Burtin, P. (1988) Immunohistochemical and biochemical study of a cathepsin B-like proteinase in human colonic cancers. *Cancer Research* **48**, 6855-6862.

Keren, Z. and LeGrue, S.J. (1988) Identification of cell surface cathepsin B-like activity on murine melanomas and fibrosarcomas: modulation by butanol extraction. *Cancer Research* **48**, 1416-1421.

Kimble, J. and Sharrock, W.J. (1983) Tissue-specific synthesis of yolk proteins in *Caenorhabditis elegans*. *Developmental Biology* **96**, 189-196.

Kimble, J. and Ward, S. (1988) Germ-line Development and Fertilization. In: *The Nematode Caenorhabditis elegans*, 191-213. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Kirschke, H. and Barrett, A.J. (1987) Chemistry of Lysosomal Proteases. In: *Lysosomes:Their Role in Protein Breakdown*, 193-238. Edited by Glaumann, H. and Ballard, F.J. London, Academic Press.

Klinkert, M.-Q., Ruppel, A., and Beck, E. (1987) Cloning of diagnostic 31/32 kilodalton antigens of *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **25**, 247-255.

Knezetic, J.A. and Felsenfeld, G. (1989) Identification and characterization of a chicken α-globin enhancer. *Molecular and Cellular Biology* **9**, 893-901.

Knox, D.P. and Kennedy, M.W. (1988) Proteinases released by the parasitic larval stages of *Ascaris suum*, and their inhibition by antibody. *Molecular and Biochemical Parasitology* **28**, 207-216.

Ko, L.J. and Engel, J.D. (1993) DNA-binding specificities of the GATA transcription factor family. *Molecular and Cellular Biology* **13**, 4011-4022.

Kominami, E. and Uchiyama, Y. (1993) Lysosomal cysteine proteinases as processing proteases: localisation in secretory granules. In: *Proteolysis and Protein Turnover*, 39-44. Edited by Bond, J.S. and Barrett, A.J. London, Portland Press.

Kornfeld, S. (1986) Trafficking of lysosomal enzymes in normal and disease states. *Journal Of Clinical Investigation* **77**, 1-6.

Kozak, M. (1983) Comparison of initiation of protein synthesis in procaryotes, eukaryotes, and organelles. *Microbiological Reviews* **47**, 1-45.

Krause, M., White Harrison, S., Xu, S.-Q., Chen, L., and Fire, A. (1994) Elements regulating cell- and stage-specific expression of the *C.elegans* MyoD family homolog *hlh-1*. *Developmental Biology* **166**, 133-148.

Kumar, S. and Pritchard, D.I. (1992) The partial characterization of proteases present in the excretory/secretoy products and exsheathing fluid of the infective (L3) larva of *Necator americanus*. *Internation Journal of Parasitology* **22**, 563-572.

Lah, T.T., Hawley, M., Rock, K.L., and Goldberg, A.L. (1995) γ-interferon causes a selective induction of the lysosomal proteases, cathepsin B and cathepsin L in macrophages . *FEBS Letters* **363**, 85-89.

Lang, L., Reitman, M., Tang, J., Roberts, R.M., and Kornfeld, S. (1984) Lysosomal enzyme phosphorylation: recognition of a protein dependent determinant allows specific phosphorylation of oligosaccharides present on lysosomal enzymes. *Journal of Biological Chemistry* **259**, 4663-4671.

Larminie, C.G.C. and Johnstone, I.L. (1996) Isolation and characterisation of four developmentally regulated cathepsin B-like cysteine protease genes from the nematode *Caenorhabditis elegans*. *DNA and Cell Biology* **15** (in press)

Lee, X., Ahmed, F.R., Hirama, T., Huber, C.P., Rose, D.R., To, R., Hasnain, S., Tam, A., and Mort, J.S. (1990) Crystallization of recombinant rat cathepsin B. *Journal of Biological Chemistry* **265**, 5950-5951.

Lewert, R.M. and Lee, C. (1954) Studies on the passage of helminth larvae through host tissues. *Journal of Infectious Disease* **95**, 13-51.

Light, A., Frater, R., Kimmel, J.R., and Smith, E.L. (1964) Current status of the structure of papain: the linear sequence, active site sulfhydryl group, and disulfide bridges. *Proceedings of the National Academy of Sciences (USA)* **52**, 1276-1283.

Lincke, C.R., Broeks, A., The, I., Plasterk, R.H.A., and Borst, P. (1993) The expression of two P-glycoprotein (*pgp*) genes in transgenic *Caenorhabditis elegans* is confined to intestinal cells. *EMBO Journal* **12**, 1615-1620.

Mach, L., Schwihla, H., Stuwe, K., Rowan, A.D., Mort, J.S., and Glossl, J. (1993) Activation of procathepsin B in human hepatoma cells: the conversion into the mature enzyme relies on the action of cathepsin-B itself. *Biochemical Journal* **293**, 437-442.

Mach, L., Mort, J.S., and Glössl, J. (1994a) Maturation of human procathepsin B. Proenzyme activation and proteolytic processing of the precursor to the mature proteinase, *in vitro*, are primarily unimolecular processes. *Journal of Biological Chemistry* **269**, 13030-13035.

Mach, L., Mort, J.S., and Glössl, J. (1994b) Noncovalent complexes between the lysosomal proteinase cathepsin B and its propeptide account for stable, extracellular, high molecular mass forms of the enzyme. *Journal of Biological Chemistry* **269**, 13036-13040.

Maciewicz, R.A., Wardale, R.J., Etherington, D.J., and Paraskeva, C. (1989) Immunodetection of cathepsins B and L present in and secreted fron human pre-malignant and malignant colorectal tumour cell lines. *International Journal of Cancer* **43**, 478-486.

MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T., and Spieth, J. (1992) Regulation of vitellogenin gene expression in transgenic *Caenorhabditis elegans:* short sequences required for activation of the *vit-2* promoter. *Molecular and Cellular Biology* **12**, 1652-1662.

Madl, J.E. and Herman, R.K. (1979) Polyploids and sex determination in *Caenorhabditis elegans. Genetics* **93**, 393-402.

Matsunaga, Y., Saibara, T., Kido, H., and Katunuma, N. (1993) Participation of cathepsin B in processing of antigen presentation to MHC class-II. *FEBS Letters* **324**, 325-330.

McCombie, W.R., Adams, M.D., Kelley, J.M., FitzGerald, M.G., Utterback, T.R., Khan, M., Dubnick, M., Kerlavage, A.R., Venter, J.C., and Fields, C. (1992) *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genetics* **1**, 124-131.

McGinty, A., Moore, M., Halton, D.W., and Walker, B. (1993) Characterization of the cysteine proteinases of the common liver fluke *Fasciola hepatica* using novel, active-site directed affinity labels. *Parasitology* **106**, 487-493.

Mello, C.C., Kramer, J.M., Stinchcomb, D., and Ambros, V. (1991) Efficient gene transfer in *C.elegans*: extrachromosomal maintenance and integration of transforming sequences. *EMBO Journal* **10**, 3959-3970.

Merika, M. and Orkin, S.H. (1993) DNA-binding specificity of GATA family transcription factors. *Molecular and Cellular Biology* **13**, 3999-4010.

Miura, M., Zhu, H., Rotello, R., Hartwieg, E.A., and Yuan, J.Y. (1993) Induction of apoptosis in fibroblasts by IL-1ß-converting enzyme, a mammalian homolog of the *C.elegans* cell death gene *ced-3. Cell* **75**, 653-660.

Moin, K., Rozhin, J., McKernan, T.B., Sanders, V.J., Fong, D., Honn, K.V., and Sloane, B.F. (1989) Enhanced levels of cathepsin B mRNA in murine tumours. *FEBS Letters* **244**, 61-64.

Morrison, R.I.G., Barrett, A.J., Dingle, J.T., and Prior, D. (1973) Cathepsins BI and D action on human cartilage proteoglycans. *Biochimica et Biophysica Acta* **302**, 411-419.

Mort, J.S., Recklies, A.D., and Poole, A.R. (1980) Characterization of a thiol proteinase secreted by malignant human breast tumours. *Biochimica et Biophysica Acta* **614**, 134-143.

Mort, J.S., Leduc, M.S., and Recklies, A.D. (1983) Characterization of a latent cysteine proteinase from ascitic fluid as a high molecular weight form of cathepsin B. *Biochimica et Biophysica Acta* **755**, 369-375.

Morton, P.A., Zacheis, M.L., Giacoletto, K.S., Manning, J.A., and Schwartz, B.D. (1995) Delivery of nascent MHC class II-invariant chain complexes to lysosomal compartments and proteolysis of invariant chain by cysteine proteases precedes peptide binding in B-lymphoblastoid cells. *Journal Of Immunology* **154**, 137-150.

Musil, D., Zucic, D., Turk, D., Engh, R.A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katanuma, N., and Bode, W. (1991) The refined 2.15Å X-ray crystal structure of human liver cathepsin B: the structural basis of its specificity. *EMBO Journal* **10**, 2321-2320.

Nishimura, Y., Kawabata, T., and Kato, K. (1988) Identification of latent procathepsins B and L in microsomal lumen: characterisation of enzymatic activation and proteolytic processing *in vitro*. *Archives of Biochemistry and Biopysics* **261**, 64-71.

Okkema, P.G., White Harrison, S., Plunger, V., Aryana, A., and Fire, A. (1993) Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135**, 385-404.

Orkin, S.H. (1992) GATA-Binding Transcription Factors in Hematopoietic Cells. *Blood* **80**, 575-581.

Otto, K. (1971) Cathepsins B1 and B2. In: *Tissue Proteinases*, 1-28. Edited by Barrett, A.J. and Dingle, J.T. Amsterdam, North-Holland Publishing Co.

Page, A.E., Warburton, M.J., Chambers, T.J., Pringle, J.A.S., and Hayman, A.R. (1992) Human osteoclastomas contain multiple forms of cathepsin B. *Biochimica et Biophysica Acta* **1116**, 57-66.

Palazzolo, M.J., Hamilton, B.A., Ding, D., Martin, C.H., Mead, D.A., Mierendorf, R.C., Raghavan, K.V., Meyerowitz, E.M., and Lipshitz, H.D. (1990) Phage lambda cDNA cloning vectors for subtractive hybridization, fusion-protein synthesis and Cre-*loxP* automatic plasmid subcloning. *Gene* **88**, 25-36.

Petrova-Skalkova, D., Krepela, E., Rasnick, D., and Vicar, J. (1987) A latent form of cathepsin B in pleural effusions. *Biochemical Medicine and Metabolic Biology* **38**, 219-227.

Pietras, R.J. and Roberts, J.A. (1981) Cathepsin B-like enzymes: subcellular distribution and properties in neoplastic and control cells from human ectocervix. *Journal of Biological Chemistry* **256**, 8536-8544.

Pohl, J., Baudys, M., Tomasek, and Kostka, V. (1982) Identification of the active site cysteine and the disulphide bonds in the N-terminal part of the molecule of bovine spleen cathepsin B. *FEBS Letters* **142**, 23-26.

Pratt, D., Cox, G.N., Milhausen, M.J., and Boisvenue, R.J. (1990) A developmentally regulated cysteine protease gene family in *Haemonchus contortus*. *Molecular and Biochemical Parasitology* **43**, 181-192.

Pratt, D., Armes, L.G., Hageman, R., Reynolds, V., Boisvenue, R.J., and Cox, G.N. (1992a) Cloning and sequence comparisons of four distinct cysteine proteases expressed by *Haemonchus contortus* adult worms. *Molecular and Biochemical Parasitology* **51**, 209-218.

Pratt, D., Boisvenue, R.J., and Cox, G.N. (1992b) Isolation of putative cysteine protease genes of *Ostertagia ostertagi*. *Molecular and Biochemical Parasitology* **56**, 39-48.

Qian, F., Bajkowski, A.S., Steiner, D.F., Chan, S.J., and Frankfater, A. (1989) Expression of five cathepsins in murine melanomas of varying metastatic potential and normal tissues. *Cancer Research* 4870-4875.

Qian, F., Frankfater, A., Chan, S.J., and Steiner, D.F. (1991) The structure of the mouse cathepsin B gene and its putative promoter. *DNA and Cell Biology* **10**, 159-168.

Ravid, K., Doi, T., Beeler, D., Kuter, D.L., and Rosenberg, R.D. (1991) Transcriptional regulation of the platelet factor 4 gene: interaction between enhancer silencer domain and the GATA site. *Molecular and Cellular Biology* **11**, 6116-6127.

Ray, C. and McKerrow, J.H. (1992) Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. *Molecular and Biochemical Parasitology* **51**, 239-249.

Recklies, A.D., Tiltman, K.J., Stoker, T.A.M., and Poole, A.R. (1980) Secretion of proteinases from malignant and nonmalignant human breast tissue. *Cancer Research* **40**, 550-556.

Recklies, A.D., Mort, J.S., and Poole, A.R. (1982a) Secretion of a thiol proteinase from mouse mammary carcinomas and its characterization. *Cancer Research* **42**, 1026-1032.

Recklies, A.D., Poole, A.R., and Mort, J.S. (1982b) A cysteine proteinase secreted from human breast tumors is immunologically related to cathepsin B. *Biochemical Journal* **207**, 633-636.

Rhaissi, H., Bechet, D., and Ferrara, M. (1993) Multiple leader sequences for mouse cathepsin B mRNA? *Biochimie* **75**, 899-904.

Richer, J.K., Sakanari, A., Frank, G.R., and Grieve, R.B. (1992) *Dirofilaria immitis:* proteases produced by third- and fourth-stage larvae. *Experimental Parasitology* **75**, 213-222.

Riddle, D.L. (1988) The Dauer Larva. In: *The Nematode Caenorhabditis elegans*, 393-412. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Rijnboutt, S., Aerts, H.M.F.G., Geuze, H.J., Tager, J.M., and Strous, G.J. (1991) Mannose-6-phosphate independent membrane association of cathepsin D, glucocerebrosidase, and sphingolipid-activating protein in HepG2 cells. *Journal of Biological Chemistry* **266**, 4862-4868.

Ritonja, A., Popovic, T., Turk, V., Wiedenmann, K., and Machleidt, W. (1985) Amino acid sequence of human liver cathepsin B. *FEBS Letters* **181**, 169-172.

Rosenfeld, M.G., Kreibich, G., Popov, D., Kato, K., and Sabatini, D.D. (1982) Biosynthesis of lysosomal hydrolases: their synthesis in bound polysomes and the role of co-translational and post-translational processing in determining their sub-cellular distribution. *Journal Of Cell Biology* **93**, 135-143.

Roussell, D.L. and Bennett, K.L. (1992) *Caenorhabditis* cDNA encodes an eIF-4A-like protein. *Nucleic Acids Research* **20**, 3783

Ruppel, A., Brekernitz, U., and Burger, R. (1987) Diagnostic M_r 31 000 *Schistosoma mansoni* proteins: requirement of infection, but not immunization, and use of the "miniblot" technique for the production of monoclonal antibodies. *Journal of Helminthology* **61**, 95-101.

Rushforth, A.M., Saari, B., and Anderson, P. (1993) Site-selected insertion of the transposon Tc1 into a *Caenorhabditis elegans* myosin light chain gene. *Molecular and Cellular Biology* **13**, 902-910.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular Cloning: A Laborarory Manual*, Second Ed., Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

San Segundo, B., Chan, S.J., and Steiner, D.F. (1985) Identification of cDNA clones encoding a precursor of rat liver cathepsin B. *Proceedings of the National Academy of Sciences (USA)* **82**, 2320-2324.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences (USA)* **74**, 5463-5467.

Sarkis, G.J., Ashcom, Hawdon, J.M., and Jacobson, J.A. (1988a) Decline in protease activities with age in the nematode *Caenorhabditis elegans. Mechanisms in Ageing and Development* **45**, 191-201.

Sarkis, G.J., Kurpiewski, M.R., Ashcom, J.D., Jen-Jacobson, L., and Jacobson, L.A. (1988b) Proteases of the nematode *Caenorhabditis elegans. Archives of Biochemistry and Biophysics* **261**, 80-90.

Schechter, I. and Berger, A. (1967) On the size of the active site in proteases. I., papain. *Biochemical and Biphysical Research Communications* **27**, 157-162.

Schechter, I. and Berger, A. (1968) On the active site of proteases. III. mapping the active site of papain; peptide inhibitors of papain. *Biochemical and Biophysical Research Communications* **32**, 898-902.

Shaw, E., Wikstrom, P., and Ruscica, J. (1983) An exploration of the primary specificity site of cathepsin B. *Archives of Biochemistry and Biophysics* **222**, 424-429.

Shaw, E. and Dean, R.T. (1980) The inhibition of macrophage protein turnover by a selective inhibitor of thiol proteases. *Biochemical Journal* **186**, 385-390.

Sheahan, K., Shuja, S., and Murnane, M.J. (1989) Cysteine protease activities and tumour development in human colorectal cancers. *Cancer Research* **49**, 3809-3814.

Sherman, F., Fink, G.R., and Hicks, J.B. (1986) *Methods in Yeast Genetics*, Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Shinagawa, T., Do, Y.S., Baxter, J.D., Carilli, C., Schilling, J., and Hsueh, W.A. (1990) Identification of an enzyme in human kidney that correctly processes prorenin. *Proceedings of the National Academy of Sciences (USA)* **87**, 1927-1931.

Sloane, B.F., Honn, K.V., Sadler, J.G., Turner, W.A., Kimpson, J.J., and Taylor, J.D. (1982) Cathepsin B activity in B16 melanoma cells: a possible marker for metastatic potential. *Cancer Research* **42**, 980-986.

Spieth, J., Denison, K., Kirtland, S., Cane, J., and Blumenthal, T. (1985) The *C.elegans* vitellogenin genes: short sequence repeats in the promoter regions and homology to the vertebrate genes. *Nucleic Acids Research* **13**, 5283-5295.

Spieth, J., Shim, Y.E., Lea, K., Conrad, R., and Blumenthal, T. (1991) *elt-1,* an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. *Molecular and Cellular Biology* **11**, 4651-4659.

Spieth, J., Brooke, G., Kuerston, S., Lea, K., and Blumenthal, T. (1993) Operons in *C.elegans*: polycistronic mRNA precursors are processed by transplicing of SL2 to downstream coding regions. *Cell* **73**, 521-532.

Steiner, D.F., Kemmler, W., Tager, H.S., and Peterson, I.D. (1974) Proteolytic processing in the biosynthesis of insulin and other proteins. *Federation Proceedings* **33**, 2105-2115.

Stinchcomb, D.T., Shaw, J.E., Carr, S.H., and Hirsh, D. (1985) Extrachromosomal DNA transformation of *Caenorhabditis elegans. Molecular and Cellular Biology* **5**, 3484-3496.

Stringham, E.G., Dixon, D.K., Jones, D., and Candido, E.P.M. (1992) Temporal and spatial expression patterns of the small heat shock (*hsp16*) genes in transgenic *Caenorhabditis elegans. Molecular Biology of the Cell* **3**, 221-223.

Stroeher, V.L., Kennedy, B.P., Millen, K.J., Schroeder, D.F., Hawkins, M.G., Goszczynski, B., and McGhee, J.D. (1994) DNA-protein interactions in the *Caenorhabditis-elegans* embryo: oocyte and embryonic factors that bind to the promoter of the gut- specific *ges-1* gene. *Developmental Biology* **163**, 367-380.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R. (1992) The *C.elegans* genome sequencing project: a beginning. *Nature* **356**, 37-41.

Sulston, J. and Hodgkin, J. (1988) Methods. In: *The nematode Caenorhabditis elegans*, 587-606. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Sulston, J.E. and Horvitz, H.R. (1977) Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology* **56**, 110-156.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Developmental Biology* 64-119.

Tabish, M., Siddiqui, Z.K., Nishikawa, K., and Siddiqui, S.S. (1995) Exclusive expression of *C.elegans osm-3* kinesin gene in chemosensory neurons open to the external environment. *Journal of Molecular Biology* **247**, 377-389.

Takahashi, K., Isemura, M., and Ikenaka, T. (1979) Isolation and characterisation of three forms of cathepsin B from porcine liver. *Journal of Biochemistry* **85**, 1053-1060.

Takahashi, K., Isemura, M., Ono, T., and Ikenaka, T. (1980) Location of the essential thiol of porcine liver cathepsin B. *Journal of Biochemistry* **87**, 347-350.

Takahashi, T., Schmidt, P.G., and Tang, J. (1984) Novel carbohydrate structures of cathepsin B from porcine spleen. *Journal of Biological Chemistry* **259**, 6059-6062.

Takio, K., Towatari, T., Katanuma, N., Teller, D.C., and Titani, K. (1983) Homology of amino acid sequences of rat liver cathepsins B and H with that of papain. *Proceedings of the National Academy of Sciences (USA)* **80**, 3666-3670.

Tam, S.W., Cote-Paulino, L.R., Peak, D.A., Sheahan, K., and Murnane, M.J. (1994) Human cathepsin B-encoding cDNAs: sequence variations in the 3'-untranslated region. *Gene* **139**, 171-176.

Taniguchi, T., Mizuochi, T., Towatari, T., Katanuma, N., and Kobata, A. (1985) Structural studies on the carbohydrate moieties of rat liver cathepsin B. *Journal of Biochemistry* **97**, 973-976.

Tata, J.R. (1976) The expression of the vitellogenin gene. *Cell* **9**, 1-14.

Tsai, S.F., Martin, D.I.K., Zon, L.I., Dandrea, A.D., Wong, G.G., and Orkin, S.H. (1989) Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446-451.

von Figura, K. and Hasilik, A. (1986) Lysosomal enzymes and their receptors. *Annual Review Of Biochemistry* **55**, 167-193.

Wall, L., Deboer, E., and Grosveld, F. (1988) The human ß-globin gene-3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes & Development* **2**, 1089-1100.

Walters, M. and Martin, D.I.K. (1992) Functional erythroid promoters created by interaction of the transcription factor GATA-1 with CACCC and AP-1/NFE-2 elements. *Proceedings of the National Academy of Sciences (USA)* **89**, 10444-10448.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., Hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierrymieg, J., and Sulston, J. (1992) A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genetics* **1**, 114-123.

Werle, B., Ebert, W., Klein, W., and Spiess, E. (1994) Cathepsin B in tumours, normal tissue and isolated cells from the human lung. *Anticancer Research* **14**, 1169-1176.

White, J. (1988) The Anatomy. In: *The Nematode Caenorhabditis elegans*, 81-122. Edited by Wood, W.B. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986) The structure of the nervous system of *Caenorhabditis elegans*. *Philosphical Transactions of the Royal Society* **314**, 1-340.

Wieringa, B., Hofer, E., and Weissmann, C. (1984) A minimal intron length but no specific internal sequence is required for splicing the large rabbit ß-globin intron. *Cell* **37**, 915-925.

Xu, M.Z., Capraro, G.A., Daibata, M., Reyes, V.E., and Humphreys, R.E. (1994) Cathepsin B cleavage and release of invariant chain from MHC class II molecules follow a staged pattern. *Molecular Immunology* **31**, 723-731.

Xue, D., Finney, M., Ruvkun, G., and Chalfie, M. (1992) Regulation of the *mec-3* gene by the *C.elegans* homeoproteins UNC-86 and MEC-3. *EMBO Journal* **11**, 4969-4979.

Yamasaki, H., Kominami, E., and Aoki, T. (1992) Immunocytochemical localisation of a cysteine protease in adult worms of the liver fluke *Fasciola* sp. *Parasitology Research* **78**, 574-580.

Yong Song, C. and Chappell, C.L. (1993) Purification and partial characterization of cysteine proteinase from S*pirometra mansoni* plerocercoids. *Journal of Parasitology* **79**, 517-524.

Young, J.M. and Hope, I.A. (1993) Molecular markers of differentiation in *Caenorhabditis elegans* obtained by promoter trapping. *Developmental Dynamics* **196**, 124-132.

Zon, L.I., Yamaguchi, Y., Yee, K., Albee, E.A., Kimura, A., Bennett, J.C., Orkin, S.H., and Ackerman, S.J. (1993) Expression of mRNA for the GATA-binding proteins in human eosinophils and basophils: potential role in gene transcription. *Blood* **81**, 3234-3241.

Zorio, D.A.R., Cheng, N.S.N., Blumenthal, T., and Spieth, J. (1994) Operons as a common form of chromosomal organization in *C.elegans. Nature* **372**, 270-272.

Zwaal, R.R., Broeks, A., Vanmeurs, J., Groenen, J.T.M., and Plasterk, R.H.A. (1993) Target-selected gene inactivation in *Caenorhabditis elegans* by using a frozen transposon insertion mutant bank. *Proceedings of the National Academy of Sciences (USA)* **90**, 7431-7435.

Appendix A

The 5' RACE experiments identified a potential splice acceptor site (TTTTCAG) in the 5' UTR of the *cpr-3* mature transcript. This suggests that the mature transcript of *cpr-3* may be trans-spliced to one of the two splice leaders, SL1 or SL2. Dr. I.Hope (pers. comm.) has reported problems amplifying SL1 or SL2 trans-spliced messages using a 5' RACE approach. Therefore, it is possible that the 5' RACE products generated for *cpr-3* may be derived from immature mRNA transcripts which are subsequently trans-spliced. This possibility should be addressed by PCR amplification of *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6* from a *C.elegans* cDNA library using the SL1 and SL2 primers and gene specific primers for each of the four genes.