Molecular characterization of a gene encoding a mucin-like protein from *Caenorhabditis elegans*.

A thesis submitted for the Degree of
Doctor of Philosophy at the
University of Glasgow

by

Sérgio Ceroni da Silva

Institute of Genetics
Church Street
Glasgow

December 1993

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

The research reported in this thesis is my own original work except where otherwise stated and has not been submitted for any other degree.

To Ângela and Kim.

LIST OF CONTENTS.

LIST OF ILLUSTRATIONS AND TABLES.

ACKNOWLEDGEMENTS.

I would like to thank my supervisor, Dave Barry, for giving me the opportunity to come to Glasgow and develop this project.

I would also like to thank everyone in the Genetics Department and at WUMP (Wellcome Unit of Molecular Parasitology), specially Iain Johnstone, for all the help and advice during this work.

Many thanks also go to the prep. room ladies, specially to Doris and Agnes, who were always very helpful.

Thanks also go to Ângela and Kim for all the love and support during the last wintry years.

Finally, I would like to thank all my friends and colleagues at Faculdade de Veterinária and Centro de Biotecnologia in Porto Alegre for all the encouragement and support to pursue my PhD studies.

This work was supported by a studentship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasil).

ABSTRACT.

A mutator strain containing a Tc1 transposon disrupting the *dpy*-6 gene activity was used to isolate the gene. This strain was backcrossed to the N2 wild-type strain in order to reduce and stabilize the Tc1 copy number. After six rounds of backcrosses, a single Tc1 insertion associated with the dumpy phenotype could be identified. This Tc1 insertion was located in a 9.5 kb *Eco*RI fragment, which was physically mapped, by hybridization to YACs and cosmids, to a region in the X chromosome of *C. elegans* were the *dpy*-6 gene is located. The Tc1 tagged fragment was cloned and the sequence adjacent to the insertion site was determined. An open reading frame containing 1857 bp was found, capable of encoding a protein with 619 amino acids. The putative protein was similar to mucin-like sequences, with the presence of thirty four copies of a six amino acid repeat arranged in tandem near the carboxy-terminus. Some of the repeats were also found dispersed in other parts of the putative protein. The amino acid repeats, like in other mucin-like proteins, were rich in threonine, serine and proline.

Another aspect which has been investigated was the identification of a collagenolytic activity in *C. elegans*. Using substrate gels containing collagen, this activity was found associated exclusively with the cuticle material. Attempts were made to isolate the corresponding gene by Polymerase Chain Reaction using oligonucleotides specific for conserved regions within metalloproteases.

ABBREVIATIONS, SYMBOLS AND MEASUREMENT UNITS.

α-^{32}P-dCTP	2'-deoxycytidine 5'-[α-^{32}P]triphosphate.
α-^{35}S-dATP	2'-deoxyadenosine 5'-[α-^{35}S]triphosphate.
ß-ME	ß-mercaptoethanol.
γ-^{32}P-ATP	adenosine 5'-[γ-^{32}P]triphosphate.
A	adenine.
A_{260}	absorbance (measured at 260 nm).
amp	ampicillin.
bp	base pair.
BSA	bovine serum albumin.
C	cytosine.
cDNA	complementary DNA.
Ci	Curie.
CIP	calf intestinal phosphatase.
dATP	2'-deoxyadenosine-5'-triphosphate.
dCTP	2'-deoxycytidine-5'-triphosphate.
dGTP	2'-deoxyguanosine-5'-triphosphate.
DNA	deoxyribonucleic acid.
DNAase	deoxyribonuclease.
DTT	dithiothreitol.
dTTP	thymidine-5'-triphosphate.
EDTA	diaminoethanetetra-acetic acid (disodium salt).
EMS	ethyl methanesulphonate.
G	guanine.
HEPES	N-[2-hydroxyethyl]piperazine-N'-[2-ethanesulfonic acid].
HSD	herring sperm DNA.

I	inosine.
IPTG	isopropyl ß-D-thiogalactopyranoside.
kb	kilobase. 1 kb is equivalent to 1000 bp.
kDa	kiloDalton.
krpm	rotations per minute multiplied by 1000.
OD	optical density.
ORF	open reading frame.
PAGE	polyacrylamide gel electrophoresis.
PCR	polymerase chain reaction.
PEG	polyethylene glycol.
pfu	plaque forming unit.
RNAase	ribonuclease.
rpm	rotations per minute.
SDS	sodium dodecyl sulphate.
ssDNA	single stranded DNA.
T	thymine.
tet	tetracycline.
tet^R	resistance to tetracycline.
TEMED	N,N,N',N'-tetramethyl-ethylenediamine.
Tris	2-amino-2-(hidroxymethyl)-1,3-propandiol.
UV	ultra-violet radiation.
V	volt.
W	watt.
X-Gal	5-bromo-4-chloro-3-indolyl-ß-D-galactopyranoside.
YAC	Yeast Artificial Chromosome.

1. INTRODUCTION.

1.1. The nematode *Caenorhabditis elegans* and its use as a model for parasitic nematodes.

C. elegans is a small (1.5 mm long) free-living nematode found in the soil of many parts of the world. It has two sexes, hermaphrodites and males. The hermaphrodites, which are chromosomally XX, produce oocytes and sperm, being capable of reproducing by self-fertilization. Males arise spontaneously at very low frequencies (less than 0.5%) and, since they are XO, it probably is the result of spontaneous X chromosome loss (Hodgkin *et al.*, 1979). The males are capable of fertilizing hermaphrodites to yield a progeny containing equal numbers of males and hermaphrodites. Hermaphrodites cannot fertilize each other.

Like other nematodes, *C. elegans* develops through five post-embryonic stages. After the oocytes are fertilized, either by the hermaphrodite's own sperm or by male sperm, the embryo develops inside an egg held in the uterus for the first few cleavages, which after approximately 3 hours is laid through the vulva. Approximately 14 hours after fertilization, the first-stage larva (L1) hatches from the egg. Then, development proceeds through three additional larval stages (L2, L3 and L4) before the animal reaches the adult stage, approximately 50 hours after fertilization (at 25°C). The four larval stages are punctuated by moults, when a new cuticle is synthesized and the old cuticle is shed.

Although the different members of the Phylum

Nematoda can have very diverse habitats and life-styles, some of them adapted to parasitism and others free-living, they all share several common morphological and anatomical features. Anatomically, nematodes are comprised of two concentric tubes: an internal tube, the intestine; and an external tube, the cuticle, hypodermis, muscle and nervous system. Separating both is the pseudocoelomic space, which also contains the gonad in the adults (Wood, 1988). This similarity is also extensive to the cuticle structure, with almost all nematodes displaying the same layered organization (Wright, 1987).

According to Wood (1988) "*C. elegans* is now probably the most completely understood metazoan in terms of anatomy, genetics, development, and behavior". From a perspective focusing on molecular biology, the availability of a detailed genetic map and a physical map, the latter represented by an ordered sets of cosmids and YACs (Yeast Artificial Chromosomes)(Coulson *et al.*, 1986, 1988), and tools like transposon tagging (Moerman *et al.*, 1986) and the generation of transgenic animals by transformation (Fire, 1986), makes *C. elegans* an incomparable system in which to address questions about nematode biology which can be directly relevant to parasitic nematodes.

More recently Sulston *et al.* (1992) have initiated a project aiming at sequencing the entire genome of *C. elegans*. Also, cDNAs have been sequenced, identifying about 1200 of the estimated 15000 genes present in *C. elegans*, more than 30% of them with significant similarity with previously identified genes from other organisms (Waterston *et al.*, 1992).

1.2. The *C. elegans* cuticle.

The cuticle of *C. elegans* is an extracellular proteinaceous structure that encloses the animal, covering the external surfaces of all hypodermal cells and also the pharynx and rectum. It is thought to serve two basic functions, both as an exoskeleton to oppose the muscles and generate force for locomotion, and as a barrier to protect the animal from its environment. The entire cuticle is shed four times during the animal development and a new cuticle is synthesized at each moult.

The cuticle is synthesized and secreted by the underlying layer of hypodermal cells. Many of these cells are multinucleate, arising by cell fusion during development. The largest, hyp-7, extends most of the length of the body. The hypodermal cells of *C. elegans* can be grouped into four main categories: the main body syncytium (hyp-7), seam cells, the hypodermal cells of the head and tail, and the interfacial hypodermal cells (White, 1988).

The main body syncytium is present at hatching and at this stage is composed of 23 nuclei, making up only the dorsal hypodermis. The ventral hypodermis is composed of 12 blast cells which divide later during development, and some of their daughters fuse to the hypodermal syncytium. This syncytium then replaces the blast cells on the ventral side. The main body syncytium is also joined by other cells derived from divisions of the seam cells. A total of 110 cells fuse and join the main hypodermal syncytium during post-embryonic development (Sulston and Horvitz, 1977).

The seam cells are organised as rows of ten cells running along each lateral line of a newly hatched L1 larva. With the exception of the most anterior, they are blast cells, dividing and contributing to the set of cells added during post-embryonic development. In the adult the seam cells fuse to form two seam syncytia, which remain separate from the main body syncytia. Seam cells are responsible for the production of the alae, which are specialized cuticle structures (longitudinal thickenings) that run along the lateral sides of the animal (Singh and Sulston, 1978).

The hypodermal cells of the head and tail are quite diverse and have been classified together based solely on their positions. The hypodermis in the head is made up of a series of six concentric annular syncytia, hyp-1 to hyp-6. The tail is composed of four hypodermal cells, hyp-8 to hyp-11. These cells are mononucleate, with the exception of hyp-10, which is binucleate.

The interfacial hypodermal cells act as interfaces between the hypodermis and organs or structures that pierce it, such as the pharynx, the anus, the vulva, the excretory duct, and many sensory receptors. These cells are often intermediate in function and morphology between the adjacent structures to which they are connected.

The hypodermal cells also have functions other than secretion of the cuticle. They have phagocytic capacities and are responsible for the elimination of cells that undergo programmed cell death during development (Robertson and Thomson, 1982). They also function as major storage cells for lipid droplets and storage granules.

The *C. elegans* cuticle is similar in structure to the cuticle of other nematodes, with at least three distinct layers being distinguished by electron microscopy: the cortical layer, the median layer, and the innermost basal layer (Cox *et al.*, 1981b). This fits the basic layered structure proposed for most nematode cuticles (Wright, 1987).

Cuticles from different larval stages of *C. elegans* exhibit remarkable differences in structure and also in their protein content (Cox *et al.*, 1981c). In the adult cuticle, the basal and cortical layers are separated by a median layer, which is probably filled with fluid. This median layer is crossed by "struts", which are dense columns of material, representing the only attachment points between the basal and cortical layers. The cortical layer can be subdivided into two distinct regions: a thin, dense (by electron microscopy) external cortical layer, and a less dense internal cortical layer. Three different regions or zones can be distinguished in the basal layer: a fibrillar layer, which is loosely organized and closer to the hypodermis; and two highly organized layers of fibres arranged in spiral around the animal in opposite directions, each at an angle of 60° to 70° relative to the body long axis.

C. elegans L4 larvae have a cuticle distinct from the adults. There is no median layer separating the basal and cortical layers. Nevertheless, the external and internal cortical layers are structurally similar. The basal layer contains two fibrillar layers, which could be organized forming oposite spirals like in the adult cuticle.

The most conspicuous characteristic of the L1 cuticle is the presence of transverse striations in the basal layer, closer

to the hypodermis. This striated zone appears to consist of intersecting sets of longitudinally and circumferentially oriented fibrils. Animals mutant at the *sqt*-3 locus have marked defects in the striated zone of the L1 cuticle and hatch as short larvae (Priess and Hirsh, 1986). The cortical layer has a structure similar to that described for the L4 and adult cuticles.

All these changes in structure are also reflected in changes in protein composition. For instance, the L1 larva cuticle has approximately double the amount of ß-ME (ß-mercaptoethanol) insoluble proteins as that found in L4 and adult cuticles (Cox et al., 1981c). Cuticles of all stages are partially solubilized by sulfhydryl-reducing agents like ß-ME. In the adult cuticle the entire basal layer, the "struts" and most of the internal cortical layer are removed by ß-ME treatment. This is also true for the L4 cuticle. In the L1 cuticle only the internal cortical layer is predominantly removed, while the external cortical layer and portions of the striated layer are insoluble. Most of these ß-ME soluble proteins are also sensitive to digestion with bacterial collagenase (Cox *et al.*, 1981b), which specifically cleaves the Y-Gly bond in the Gly-X-Y repeats of the collagen primary structure (X and Y can be any amino acid, but often are either proline or hydroxyproline; Fessler and Fessler, 1978). It has also been shown that the synthesis of cuticle components, both collagenous and non-collagenous, occurs at high levels during the moulting periods and at a much reduced rate during the intermoult periods (Cox *et al.*, 1981a).

The remarkable changes in cuticle structure and composition occurring during *C. elegans* development have led Cox *et al.* (1981c) to suggest that they might be an evolutionary

vestige of a parasitic ancestry, since stage-specific differences in cuticle morphology are common among parasitic nematodes and in some cases are clearly of adaptative significance to particular environments or hosts.

Since the most abundant proteins present in the cuticle are collagens, much attention has been concentrated on the elucidation of the structure and expression of the collagen genes in *C. elegans*. The first collagen genes identified from *C. elegans* were *col*-1 and *col*-2, isolated using a chicken collagen cDNA probe (Kramer *et al.*, 1982). Based on genomic Southern blot hybridization experiments and recombinant phage library screenings, using these genes as probes under low stringency conditions, so that cross hybridization between different collagen genes could occur, at least 40 and as many as 150 distinct collagen genes were detected (Cox *et al.*, 1984). These collagen genes are not organized in tandem arrays, but are dispersed in the *C. elegans* genome.

Several other collagen genes from *C. elegans* have also been cloned (Cox *et al.*, 1989), all with a similar, but not identical, structure. They are less than 1.5 kb in length, encoding proteins of 28 to 32 kDa. Approximately one-half of the protein is composed of the Gly-X-Y repeats, which contain several short interruptions of 2 to 21 amino acids. Most of the cysteine residues are found outside the repeats, but a few can be found within repeats. Based upon amino acid sequence similarities (size and position of the repeats and cysteine residues) the genes were placed into three families. Crosslinking of collagen molecules occurs primarily at the cysteine residues, via disulfide bridges (Politz *et al.*, 1986), but other non-reducible

crosslinks were also suggested to be present (Fetterer and Rhoads, 1990).

The *C. elegans* collagen genes differ significantly from the structure of vertebrate fibrillar collagen genes. The *C. elegans* genes contain only one or two small introns, whereas the vertebrate collagen genes are interrupted by more than fifty introns. Also, the vertebrate genes are quite large, approximately 50 kb, encoding proteins of 100 to 200 kDa which have uninterrupted Gly-X-Y repeats (Bornstein and Sage, 1980; Ohkubo *et al.*, 1980; Wozney *et al.*, 1981; Boedtker *et al.*, 1983).

When total RNA was analysed by Northern blot, using *col-1* or *col-2* as hybridization probes, most of the collagen bands detected were 1.2 to 1.4 kb in length. However, bands corresponding to sizes of 3 and 4 kb were also detected in much smaller amounts, presumably representing large, vertebrate-like fibrillar or basement membrane type collagens (Cox *et al.*, 1984).

It is likely that the majority of collagen genes in *C. elegans* are involved in cuticle formation. Collagen mRNAs are most abundant in the hypodermal cells (Edwards and Wood, 1983), which secrete the cuticle, and transcripts from these genes are very abundant at moults, when the cuticle is synthesized (Cox and Hirsh, 1985).

The expression of the *C. elegans* collagen genes is regulated during development. For instance, the abundance of *col-1* and *col-2* transcripts varies during the life cycle, and the expression patterns for the two genes are different. On the one hand, *col-1* transcripts are abundant during the formation of the L1 and dauer cuticles and at much lower levels during L4 and adult cuticle synthesis. On the other hand, *col-2* transcripts can

only be detected at the time when dauer cuticle synthesis is occurring (Kramer *et al.*, 1985).

Cox and Hirsh (1985) analysed the temporal patterns of mRNA accumulation for a large number of collagen genes by screening recombinant phages and plasmids containing cloned collagen genes under high stringency conditions with ^{32}P-labelled cDNA preparations specific for eggs or three post-embryonic moults (L2s moulting into dauer larvae, dauer larvae moulting into L4s, and L4s - from dauers - moulting into adults). They found that collagen mRNA levels were regulated both temporally and quantitatively during *C. elegans* development, and that most of the genes studied exhibited one of four patterns of mRNA accumulation which correlated with changes in cuticle morphology and collagen protein composition during development. These four gene classes were: 1) class A, composed of genes whose mRNAs were present at various levels in eggs and at each of the three post-embryonic moults analysed; 2) class B gene mRNAs were present at each moult but appeared to be absent from eggs; 3) class C gene mRNAs were detected only at the moults entering the adult and dauer larvae stages; and 3) class D gene mRNAs were highly abundant in animals moulting into dauer larvae but were present at much reduced levels at subsequent moults. Generally, there was a progressive activation of new collagen genes during normal development. Cox and Hirsh (1985) proposed that once a gene is activated it remained active during subsequent moults, although its level of expression could change.

The collagen genes of some parasitic nematodes have also been characterized and revealed similarities to those of

C. elegans. In *Haemonchus contortus*, a parasitic nematode of sheep, several genes and partial cDNAs encoding putative cuticle collagens have been isolated (Shamansky *et al.*, 1989). Genomic Southern blot analysis suggested that at least 20 collagen genes were present. These genes were small, less than 3 kb in length, and the different gene family members had related, but not identical primary sequences. All *H. contortus* genes analysed had a striking homology to the *C. elegans* subfamily of collagen genes represented by *col*-1. As in *C. elegans*, the *H. contortus* collagen mRNAs were small, being between 1 kb and 1.2 kb. The analyses also suggested that, similarly to the situation in *C. elegans*, the collagen genes of *H. contortus* could be dispersed in the genome.

Kingston *et al.* (1989) cloned and characterized two partial sequences of collagen genes from *Ascaris suum*, one of them containing almost the entire gene and the other only the 3' end. Once more, these genes showed a high degree of similarity to the *C. elegans col*-1 collagen gene, with almost identical size and positions of the Gly-X-Y domains. The location of some of the cysteine residues was also conserved. However, the size and position of introns in *A. suum* differed. The introns present in *A. suum* were much larger than those in *C. elegans*. One of the introns, interrupting the Gly-X-Y domain closer to the 3' end of the gene, was 560 bp in length. Northern blot analysis showed collagen mRNAs of 1.1 kb and 1.4 kb in length. Using poly(A)$^+$ RNA extracted from specific tissues in dot blot experiments, it was possible to demonstrate that one of the genes was expressed exclusively in the adult body wall tissue but not in the ovarian or gut tissue, suggesting that it encoded a cuticular collagen.

Another class of cuticular proteins present in

C. elegans is the cuticlins. These proteins, which were first described in the cuticle of *Ascaris* (Fujimoto and Kanaya, 1973), are chemically distinct from collagens and are extremely insoluble, even in detergents and chaotopric agents, due to non-reducible covalent crosslinks.

Recently, Sebastiano *et al.* (1991) identified a new gene of *C. elegans* encoding an insoluble component of the cuticle, cuticlin 1. Transcripts of this gene, which mapped to a region on chromosome II of *C. elegans* between *unc*-4 and *sqt*-1, were detected by Northern blot analysis only in mRNAs prepared from animals entering the dauer pathway. Using immunofluorescence to localize the product of the gene *cut*-1 confirmed that it was restricted to the cuticle of the dauer larvae and was not found in the cuticle of any other stage. In the dauer larvae it formed two ribbons 1 to 2 μm wide running along the lateral lines underneath the alae, and was part of a fraction of the cuticular proteins that were not solubilized by SDS and ß-ME. However, the protein could be found in worm extracts in a soluble form before it was assembled in the cuticle. Once that took place, the protein became insoluble. The location of the protein in the cuticle suggested that its function could be related to the strengthening of the seam between the two halves of the cuticle. The gene was composed of four exons, 118 bp, 227 bp, 736 bp, and 340 bp in length, capable of encoding a predicted protein of 423 amino acids.

Another cuticlin gene, *cut*-2, was expressed in all developmental stages of *C. elegans* and the predicted protein contained only 283 amino acids. This protein contained 12 repeat motifs including one tyrosine residue per repeat. Four

copies of this tyrosine-containing motif were also found in *cut*-1, suggesting that it could be involved in the crosslinking of cuticlin units, through dityrosine bridges. The *cut*-2 gene product was present in the external cortical layer of all stages, but not on the surface of the cuticle (cited in Politz and Philipp, 1992).

1.3. Genes affecting the structure of the cuticle and body morphology.

Amongst the best characterized genes with conspicuous effects on body morphology are the dumpy genes. These genes were first identified using mutagenesis induced by EMS (ethyl methanesulphonate), and were mapped to all chromosomes of *C. elegans* (Brenner, 1974). Dumpy animals are shorter than the wild-type, but with the same diameter, giving the animal a "fat", chunky appearance. Currently, 28 different dumpy genes have been identified (Wood, 1988).

The first dumpy gene from *C. elegans* to be cloned and characterized was the *dpy*-13 gene (von Mende et al., 1988). The gene was isolated by transposon tagging, from a mutator strain with a high frequency of Tc1 transposition. Sequencing revealed that the *dpy*-13 gene encoded a collagen protein. The *dpy*-13 gene contained four domains potentially encoding Gly-X-Y repeats. Comparison with other collagen sequences revealed a great similarity to the collagen genes *col*-1 and *col*-2, not only in the size and location of the repeats but also in the flanking sequences corresponding to the amino and carboxy termini of

the protein. The location of some cysteine residues were also conserved. *dpy*-13 is interrupted by two short introns of 58 bp and 50 bp, respectively. Apparently, because this gene yields a mutant phenotype when mutated singly, its function is not replaced by other collagens of related structure, and it might have a unique function within the collagen gene family.

Another dumpy gene from *C. elegans*, *dpy*-7, has been cloned. Using the collagen gene *col*-1 as a probe to screen the DNA prepared from YACs spanning the region of the X chromosome were *dpy*-7 had previously been mapped, Johnstone *et al.* (1992) identified a homologous collagen sequence. The *dpy*-7 gene was structurally similar to other *C. elegans* collagen genes. However, it did not conform to any of the previously described family of collagen genes, as defined by the structure of Gly-X-Y repeats and location of cysteine residues (Cox *et al.*, 1989). Four EMS-induced mutant alleles of *dpy*-7 were also analysed and all of them contained G to A transitions affecting glycine residues within the Gly-X-Y repeats, resulting in glycine to arginine substitutions in three of them and a glycine to aspartate substitution in another allele analysed. These findings indicated that changes in the glycine residues present in the repeats were disruptive to the assembling of the triple helix structure of collagen.

Two other dumpy genes from *C. elegans*, *dpy*-2 and *dpy*-10, have recently been cloned (Levy *et al.*, 1993). Mutations in these genes have previously been demonstrated to cause a variety of phenotypes, among them animals with dumpy, left roller (animals twisted in a left-handed helix, which rotate around their long axis and tend to move in circles), and

dumpy/left roller phenotypes (Brenner, 1974; Cox *et al.*, 1985; Kush and Edgar, 1986). The *dpy*-2 and *dpy*-10 genes, which encoded collagens, were only 3.5 kb apart and their encoded proteins had a high degree of similarity, with identical patterns of conserved cysteines, indicative that these genes were the result of a gene duplication event. Despite the overall similarity to other *C. elegans* collagen genes, they probably represented a novel subfamily. Even though both genes had a high degree of similarity, they did not appear to be functionally redundant, since a *dpy*-10 null mutant was not rescued by the *dpy*-2 gene. The nucleotide sequence of several mutants was determined in order to establish the nature of the molecular defects causing the morphological phenotypes. It was found that different glycine substitutions within the Gly-X-Y repeats were responsible for the dumpy, dumpy/left roller, or temperature-sensitive dumpy/left roller phenotypes. In one temperature-sensitive dumpy/left roller allele of *dpy*-10 an arginine to cysteine substitution was found, outside the Gly-X-Y repeats, closer to the amino-terminus of the protein. Another *dpy*-10 allele, causing a dumpy/left roller phenotype, had a non-sense codon near the end of the Gly-X-Y region, suggesting that this was the null phenotype for the gene.

Another *C. elegans* gene which can be mutated to cause a variety of morphological abnormalities is *sqt*-1 (Cox *et al.*, 1980; Kush and Edgar, 1986). Different alleles of this gene can produce animals with dumpy, long, right roller or left roller phenotypes. Using an experimental approach including physical mapping of chromosomal deficiencies, heterologous probing of cosmids with a *col*-2 probe and Tc1 transposon tagging, the *sqt*-1 gene has been cloned and determined to encode a collagen

polypeptide (Kramer *et al.*, 1988). Structurally, the *sqt*-1 gene is very similar to other *C. elegans* collagen genes. An interesting finding about *sqt*-1 was that the null phenotype of the gene was essentially wild-type, suggesting that its product is non-essential or replaceable by another similar, though not identical, collagen. *sqt*-1 has a high number of dominant alleles, a common attribute of genes encoding proteins that form multimeric structures. This suggests that mutant *sqt*-1 products could be incorporated into the cuticle, disturbing its normal structure.

It has been found that *sqt*-1 interacts genetically with some other genes affecting body morphology in *C. elegans*, including *dpy*-10, *dpy*-12, *rol*-6, *rol*-8, and *sqt*-2 (Cox *et al.*, 1980; Kush and Edgar, 1986). For instance, in the interaction with *rol*-6, a recessive left roller allele of *sqt*-1 can function as a dominant suppressor of a recessive right roller allele of *rol*-6, so that an animal homozygous for *rol*-6 and heterozygous for *sqt*-1 has a wild-type phenotype. Two classes of *rol*-6 roller allele have been identified, recessive and semi-dominant right roller alleles. The semi-dominant alleles display a reduced degree of helical twisting in the heterozygous state. The null phenotype of *rol*-6, like that of *sqt*-1, is wild-type.

The *rol*-6 has been cloned by Kramer *et al.* (1990) and shown to encode a collagen polypeptide. Its deduced amino acid sequence is more similar to the *sqt*-1 collagen than to any other *C. elegans* collagen sequence. It is particularly significant that the locations of the cysteine residues flanking the Gly-X-Y repeats of *rol*-6 and *sqt*-1 are identical but different from those of the other sequenced collagens. Also, the carboxy-terminal domains of both genes are identical in length, differing from all

other collagens, suggesting they are members of a new subfamily of collagens. The amino acid sequence similarity supports the notion that they may physically interact via their carboxy-terminal domains into heterotrimeric molecules.

Considering the high number of collagen genes in *C. elegans* and that many of them are probably involved in the assembly of the cuticle, it is possible that many dumpy phenotypes are caused by mutations in these collagen genes. However, not all *dpy* genes are likely to code for collagens. In principle, any gene with effect on the synthesis, processing or secretion of collagens or other cuticular proteins could be detected as a dumpy mutant.

Two dumpy genes, *dpy*-21 and *dpy*-26, are unusual in that despite being located on autosomes they have phenotypic consequences that are dependent on X chromosome dosage (Hodgkin, 1983). For instance, at high X chromosome to autosome ratio (animals 2A, 3X), these dumpy mutations are lethal; at intermediate ratio (animals 2A, 2X - hermaphrodites), they cause a dumpy phenotype or lethality; and at low ratio (animals 2A, 1X - males), they cause neither a dumpy phenotype nor lethality. It has been proposed that these genes are involved in an X chromosome dosage compensation mechanism, operating to prevent hyper-expression of X chromosomes in hermaphrodites. In males, some mechanism would block the function of these genes, permitting the single X chromosome to be expressed normally. Mutations in two other genes, *dpy*-27 and *dpy*-28, appeared to have similar effects to those of *dpy*-26, but with a slightly lower lethality (Hodgkin, 1988). Possibly, the dumpy phenotype in these cases could be an indirect effect of

altered collagen ratios resulting from overexpression of one or more X-linked genes.

In summary, the *C. elegans* cuticle is a very complex. multilayered extracellular structure where the concerted expression and assembly of several collagen polypeptides interact to give its physical characteristics, with other non-collagenous proteins possibly playing a role.

1.4. Aims of the present work.

The main aim of this work is to clone and characterize the *dpy*-6 gene, isolated by Tc1 transposon tagging in a mutator strain. Elucidation of the identity of this gene would contribute to the understanding of cuticle structure and factors determining morphological alterations in *C. elegans*.

A secondary aim is to determine whether a collagenolytic activity can be found associated with the cuticle material of *C. elegans* and to attempt to clone the corresponding gene. If such an activity can be detected and the corresponding gene cloned, this would be the starting point in determining whether enzymatic degradation of the cuticle assists in the moulting process.

2. MATERIALS AND METHODS.

2.1. Enzymes and reagents.

Restriction enzymes, DNA polymerase I (Klenow fragment), T4 DNA ligase, Taq DNA polymerase and T4 polynucleotide kinase were purchased from GIBCO-BRL or Promega. CIP (Calf Intestinal Phosphatase) was purchased from Boehringer Mannhein GmBH.

Proteinase K, DNAase (deoxyribonuclease I, type I, from bovine pancreas) and RNAase (ribonuclease A, from bovine pancreas) were purchased from Sigma Chemical Company.

Enzymes were used with buffers supplied or recommended by the manufacturers, unless otherwise stated.

Reagents used were either molecular biology grade or of similar quality.

2.2. Bacteria, vectors and plasmids.

Bacterial strains used during this work were all *Escherichia coli* derivatives:

1) the strain OP50 (Brenner, 1974) was routinely used for feeding the nematodes.

2) the strain XL1-Blue[1] (Bullock *et al.*, 1987) was used as host for growing λZAP II and pBluescript II vectors. It allows

---

[1] genotype: *rec*A1, *end*A1, *gyr*A96, *thi*-1, *hsd*R17 (rk-, mk+), *sup*E44, *rel*A1, *lac* [F', *pro*AB, *lac*I^{q} ZΔM15, Tn10 (tet^{R})].

blue/white colour selection when used with these vectors and plates containing X-Gal and IPTG.

The vectors used were:

1) λZAP II (Short *et al.*, 1988) is a phage vector primarily devised for cloning cDNAs up to 10 kb in length, with a polylinker containing six unique cloning sites.

2) pBluescript II (Alting-Mees and Short, 1989) is a phagemid vector containing 21 unique cloning sites flanked by six different primer sites for DNA sequencing. This plasmid has the f1 origin of replication, allowing rescue of ssDNA for sequencing.

3) pPD21.28 (Fire *et al.*, 1990) is a plasmid vector containing the *E. coli* ß-galactosidase encoding region (the *lacZ* gene) preceded by a polylinker with seven unique cloning sites, a synthetic intron segment and the SV40 nuclear localization signal. This plasmid allows the construction of fusion proteins that can be expressed in *C. elegans* cells and their location identified by staining with X-Gal.

Other plasmids used were:

1) pRF4 (Kramer *et al.*, 1990) carries a 4 kb *Eco*RI fragment of *C. elegans* genomic DNA containing the *rol*-6(su1006) collagen gene in the Bluescribe[2] vector. Animals carrying this mutant collagen gene exhibit a helically twisted cuticle and body. It was used as a dominant genetic marker for DNA transformation in *C. elegans*.

2) 6p6.1 (Lepage and Gache, 1990) is the pBluescript plasmid with a 2.5 kb *Eco*RI fragment containing almost the entire sea urchin collagenase-like hatching enzyme coding sequence.

---

[2] Stratagene.

3) pCe1003 (Emmons and Yesner, 1984) carries a 2.65 kb *Hind*III fragment of *C. elegans* DNA containing the Tc1 transposon cloned into pBR322. This plasmid was used for preparing Tc1 specific probes.

2.3. Nematode strains.

The *Caenorhabditis elegans* wild-type Bristol strain N2 (Brenner, 1974) was obtained from the Cambridge collection. The strain CB14 [*dpy*-6(e14)] (Brenner, 1974) was obtained from the Caenorhabditis Genetic Center. The strain DR1013 [*dpy*-6(m445), *mut*-6(st702), *unc*-22(st192, st527)], which is a mutator strain derived from Bristol/Bergerac hybrids (Moerman and Waterston, 1984), was a gift from P. Albert and D. Riddle, University of Missouri.

2.4. Growth and maintenance of *C. elegans*.

C. elegans was grown in small (55 mm diameter) NGM[3] plates seeded with *Escherichia coli* strain OP50 and kept at 20°C, as described by Brenner (1974). When necessary to grow large quantities of nematodes (as for DNA preparations), plates of 90 mm diameter were used and bacteria added as follows. Single

---

[3] NGM contains 0.3 g% NaCl, 0.25 g% peptone, 0.5 mg% cholesterol, 1 mM $CaCl_2$, 1 mM $MgSO_4$, 25 mM potassium phosphate (pH 6) and 1.7g% agar (or agarose, when used for DNA preparations).

colonies of OP50 were inoculated into 400 ml of L-broth[4] and incubated at 37°C for 14-16 h, after which the bacteria were harvested by centrifugation and resuspended in 10-20 ml of M9 buffer[5]. Approximately 1-2 ml from this bacterial suspension were added to each worm-containing plate every 2-3 days and plates incubated for as long as 2 weeks, when almost all bacteria had been consumed.

2.5. Genetic methods.

2.5.1. Self-fertilization.

In order to isolate progeny derived from a single worm, an L4 hermaphrodite was transferred to a seeded NGM plate and incubated at 20°C. Since hermaphrodites before the adult phase are incapable of mating, all the progeny arising from this nematode did so from self-fertilization.

2.5.2. Crossing.

Crossings were done by transferring a single L4 hermaphrodite together with 10-15 L4 males to a small NGM plate, followed by incubation at 20°C for 2-3 days. Under these conditions most of the animals were produced by self-

---

[4] L-broth contains 1% tryptone, 0.5% yeast extract, 0.5% NaCl; pH 7.2

[5] M9 buffer contains 0.3 g% KH_2PO4, 0.6 g% Na_2HPO4, 0.5 g% NaCl, 1 mM $MgSO_4$.

fertilization and genetic markers, when necessary, were used in order to distinguish between self-progeny and animals resulting from crossing.

2.6. Extraction and purification of *C. elegans* DNA.

Nematodes were grown in 5 large (90 mm diameter) NGM-agarose plates and harvested by washing each plate with 5 ml of M9 medium. The nematode suspension was transferred to a 50 ml Falcon tube, centrifuged[6] and the supernatant discarded. Nematodes were then resuspended in 30 ml of M9 and the last procedure repeated twice.

DNA was extracted essentially as described by Sulston and Hodgkin (1988). Nematodes were resuspended in 20 ml of lysis buffer[7] and incubated at 50°C for 2 h with occasional, gentle mixing. Samples were then extracted with phenol (15 minutes on a roller at minimum speed), centrifuged[8], the aqueous phase was removed and the extraction repeated with chloroform. DNA was precipitated by adding 30 ml of ethanol, spooled in a glass rod, washed in 70% ethanol, air-dried and resuspended in 1 ml of TE[9].

---

[6] 3.5 krpm for 15 minutes (4°C), rotor TH-4, TJ-6 Beckman centrifuge.

[7] lysis buffer contains 100 mM NaCl, 100 mM Tris-Cl (pH 8), 50 mM EDTA, 1% SDS, 1% ß-ME, 100 µg/ml proteinase K.

[8] 10 krpm for 15 minutes (4°C), rotor JA-20, J2-21 Beckman centrifuge.

[9] TE contains 10 mM Tris-Cl, 1 mM EDTA; pH 8.

DNA was further purified by centrifugation in a CsCl/ethidium bromide density gradient. Samples were diluted down to 8 ml with TE, 8 g of CsCl and 800 µl of ethidium bromide (10 mg/ml) were added, samples loaded in a QuickSeal™ tube and centrifuged at 40 krpm for 65 h (20°C)[10]. The band of chromosomal DNA was collected with a syringe fitted with a wide-bore needle and extracted three to four times with 1 volume of 1-butanol. CsCl was removed by centrifugation in a Centricon-30 microconcentrator (Amicon), samples resuspended in 1 ml TE (plus 0.1 M NaCl) and DNA precipitated with 2 volumes of ethanol. Finally, samples were centrifuged[11], the supernatant discarded and the DNA was resuspended in 100 µl TE.

2.7. Purification of phage lambda DNA.

Phage lambda was prepared by infecting the appropriate *E. coli* host in liquid culture. A single bacterial colony was inoculated into 3 ml of L-broth (supplemented with 10 mM $MgSO_4$, 0.2% maltose) and incubated at 37°C for 14-16 h. A 2 ml sample was withdrawn, inoculated into 200 ml of L-broth (in a 1 l flask) and incubated at 37°C, with vigorous agitation, until the culture reached mid-log phase (approximately 2 h). The cultures were then inoculated with $1\text{-}5x10^9$ pfu of phage lambda and incubation continued for

---

[10] rotor Ti50.1, L8-55 Beckman ultracentrifuge, slow deceleration.

[11] 15 krpm for 15 minutes, Heraeus Biofuge 15 centrifuge.

5-6 h. 10 ml of chloroform were added and the culture was incubated at 37°C with vigorous shaking for 15 minutes. The culture was then transferred to a large (250 ml) centrifuge flask (avoiding carrying over any chloroform) and treated with DNAase and RNAase (final concentrations of 1 μg/ml) for 30 minutes at room temperature. Solid NaCl was added (1 M final concentration), completely dissolved and samples incubated on ice for 1 h. Any cell debris were cleared by centrifugation at 10 krpm for 10 minutes (4°C)[12]. In order to precipitate the phage particles, PEG-8000 was dissolved to a final concentration of 10% (w/v) followed by incubation on ice for another 1 h. Precipitated phage particles were collected by centrifugation at 10 krpm for 15 minutes (4°C) and resuspended in 2 ml of SM buffer[13].

Phage were purified in a CsCl step gradient as described by Thomas and Davis (1975). The step gradients were prepared in 5 ml tubes[14] by carefully layering 1 ml of CsCl solutions[15] of decreasing densities (1.6, 1.5, 1.4 and 1.3 g/ml, from bottom to top). The phage suspension was then layered on the top and the tubes were centrifuged at 35 krpm for 1 h at 4°C[16]. A white band of phage could then be visualized in the

---

[12] rotor JA-14, J2-21 Beckman centrifuge.

[13] SM buffer contains 100 mM NaCl, 8.1 mM $MgSO_4$, 50 mM Tris-Cl (pH 7.5), 0.01% gelatin.

[14] Beckman, Ultra-Clear™.

[15] dissolved in 10 mM Tris-Cl (pH 7.5), 10 mM $MgSO_4$.

[16] rotor SW50.1, L8-55 Beckman ultracentrifuge.

interface between the 1.4 and 1.5 solutions and collected with a capillary tube or Pasteur pipette. CsCl was removed by centrifugation through a Centricon-30 microconcentrator. Phage were resuspended in 500 µl of SM buffer and treated with EDTA/proteinase K/SDS[17] at 56°C for 1 h. The samples were then extracted successively with phenol, phenol/chloroform (1:1) and chloroform. DNA was precipitated by adding NaCl, to a final concentration of 0.1 M, and 2 volumes of ethanol, collected by centrifugation and resuspended in 100 µl TE.

2.8. Preparation of plasmid and cosmid DNA.

Plasmid and cosmid DNA were purified by a modification of the alkaline-lysis method described by Birnboim and Doly (1979). Approximately 1.5 ml of an overnight culture was centrifuged[18] and all the supernatant removed with a capillary. The bacterial pellet was resuspended in 100 µl of solution I[19] and incubated for 5 minutes at room temperature, 200 µl lysis solution[20] was then added and the preparations were incubated on ice for 5 minutes. The bulk of chromosomal DNA and proteins were precipitated by adding 150 µl of 3 M sodium

---

[17] final concentrations of 20 mM, 50 µg/ml and 0.5%, respectively.

[18] 10 krpm for 5 minutes, Heraeus Biofuge 15.

[19] solution I contains 50 mM glucose, 25 mM Tris-Cl (pH 8), 10 mM EDTA.

[20] 0.2 N NaOH, 1% SDS; freshly made.

acetate (pH 5.2)[21] followed by centrifugation[22]. The supernatant was transferred to a clean Eppendorf tube, treated with RNAase[23] and DNA precipitated with 0.6 volumes of isopropanol. Finally, DNA was collected by centrifugation and resuspended in 20-100 µl of TE.

The same method was used for the preparation of larger quantities of DNA, scaled up for starting cultures of up to 45 ml. The only difference was additional extractions with phenol and phenol/chloroform (1:1) immediately before the isopropanol precipitation.

2.9. Synthesis and purification of oligonucleotides.

Oligonucleotides were synthesized in a PCR-mate (Applied Biosystems 391 DNA synthesizer) using Cruachem reagents. After synthesis was completed, the matrix was removed from the column and dissolved in 1 ml 100% ammonium hydroxide and incubated for 2 h at room temperature. The matrix was sedimented by centrifugation[24] and the supernatant transferred to clean screw-cap Eppendorf tubes. A total of 2 ml ammonium hydroxide was added and the samples were incubated overnight at 55°C.

---

[21] pH 4.8 when used for cosmid preparations.

[22] 15 krpm for 15 minutes, Heraeus Biofuge 15.

[23] 50 µg/ml final concentration, for 30 minutes at 37°C.

[24] 15 krpm for 5 minutes, Heraeus Biofuge 15.

Oligonucleotides were precipitated by adding 1/10 volume 5 M ammonium acetate and 2.5 volumes of ethanol and recovered by centrifugation[25]. Following this, the supernatant was discarded and the DNA pellet dried and resuspended in 200 µl TE.

Concentration was estimated by measuring A_{260} in a spectrophotometer[26] and using the extinction coefficients given by Sambrook *et al.* (1989).

2.10. Gel electrophoresis of DNA.

DNA fragments were separated by agarose gel electrophoresis using TBE buffer[27]. Different agarose concentrations were used, ranging from 0.4% to 2%, depending on the fragment sizes to be resolved. The electrophoresis system used was the horizontal with gel submerged in buffer, essentially as described in Sambrook *et al* (1989). 1/10 volume sample buffer[28] was added to DNA prior to loading the gel. All samples running in a single gel had their ionic strength corrected to the highest amongst the samples. When maximum resolution of DNA fragments was to be achieved, gels were electrophoresed by

---

[25] 15 krpm for 10 minutes, Heraeus Biofuge 15.

[26] Beckman, model DU-50.

[27] 1X TBE contains 89 mM Tris-OH, 2.5 mM EDTA, 89 mM H_3BO_3 pH 8.2.

[28] 25% Ficoll™ 400000, 0.25% bromophenol blue, 0.25% xylene cyanol.

applying voltage gradients of 2 V/cm[29] (corresponding to a running time of 14-18 h). Otherwise, gels were routinely run using voltage gradients of 10 V/cm (running time of approximately 1 h). Gels were stained in ethidium bromide to a final concentration of 0.5 μg/ml (Sharp *et al.*, 1973) and photographed under UV transillumination, using a Polaroid MP-4 Land camera (type 667 film).

DNA fragment sizes were estimated from these photographs using a computer programme based on the Schaffer and Sederoff (1981) algorithm. Molecular weight standards used were phage Lambda DNA *Hind*III digested (λ/*Hind*III), phage ΦX-174 DNA *Hae*III digested (ΦX-174/*Hae*III) or a 1 kb ladder[30].

2.11. Isolation and purification of DNA fragments separated by agarose gel electrophoresis.

After digestion with the appropriate restriction enzymes DNA fragments were separated by electrophoresis in LMP-agarose[31] gels which were then stained with ethidium bromide and viewed under long-UV (300 nm) transillumination. The relevant bands were excised from the gel, then the agarose was melted at 65°C for 10-15 minutes and diluted with 5 volumes of 20 mM Tris-Cl (pH 8), 1 mM EDTA. Agarose was

---

[29] in relation to gel length.

[30] GIBCO-BRL.

[31] low melting point agarose (GIBCO-BRL).

removed by successive extractions with phenol and phenol/chloroform (1:1) and DNA was precipitated by adding 1/3 volume of 10 M ammonium acetate and 2 volumes of ethanol. DNA was collected by centrifugation[32] and resuspended in 50-200 µl of TE.

Alternatively, DNA fragments were recovered from conventional agarose gels using the freeze-squeeze method (Tautz and Renz, 1983).

2.12. DNA blotting.

DNA was transferred from agarose gels to nylon membranes[33] using the technique described by Southern (1975).

Immediately after electrophoresis, the gel was immersed in denaturing solution[34] for 30 minutes at room temperature with agitation on a rocking platform. The denaturing solution was discarded and the gel was immersed in neutralizing solution[35] for 45 minutes. The gel was then placed on wetted 3 MM paper mounted in a receptacle containing 20X SSC[36]. A nylon membrane was placed on top of the gel and two layers of 3 MM paper over the membrane. Several paper towels

---

[32] 10 krpm for 15 minutes, rotor JA-20, J2-21 Beckman centrifuge.

[33] Hybond-N, Amersham.

[34] denaturing solution contains 1.5 M NaCl, 0.5 M NaOH.

[35] neutralizing solution contains 1.5 M NaCl, 0.5 M Tris-Cl; pH 7.

[36] 1X SSC contains 150 mM NaCl, 15 mM tri-sodium citrate.

were stacked over this assembly and capillary transfer allowed to proceed for 14-16 h, after which the nylon membrane was baked at 80°C for 2 h.

2.13. End-labelling of oligonucleotides.

Synthetic oligonucleotides were labelled at their 5' termini in reactions containing 3 pmol of the required oligonucleotide, 50 mM Tris-Cl (pH 7.6), 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM spermidine HCl, 0.1 mM EDTA (pH 8), 140 µCi γ-^{32}P-ATP (6000 Ci/mmole) and 10 U of T4 polynucleotide kinase, in a final volume of 20 µl. Reactions were incubated at 37°C for 45 minutes and then heat inactivated at 68°C for 10 minutes. Samples were diluted with 73 µl H_2O and precipitated by the addition of 107 µl 7.5 M ammonium acetate and 500 µl ethanol. Finally, samples were incubated at -20°C for 1-2 h and oligonucleotides recovered by centrifugation[37] and resuspended in 100 µl TE.

2.14. DNA labelling.

DNA was labelled by the random priming technique (Feinberg and Vogelstein, 1983, 1984) using α-^{32}P-dCTP.

Labelling reactions contained 10-50 ng of DNA, 60 mM Tris-Cl (pH 8), 6 mM $MgCl_2$, 12.5 mM ß-ME, 24 µM dGTP, 24 µM

---

[37] 15 krpm for 30 minutes (4°C), Heraeus Biofuge 15.

dATP, 24 μM dTTP, 0.25 M HEPES (pH 6.6), 6.7 OD units/ml random hexadeoxyribonucleotides (Pharmacia), 0.5 mg/ml BSA, 30 μCi α-^{32}P-dCTP (3000 Ci/mmole) and 1-2 units of Klenow DNA polymerase, in a final volume of 40 μl. Firstly, DNA samples were denatured by boiling for 5 minutes in a water bath and then they were cooled rapidly on ice. After adding the other components, reactions were incubated at room temperature for 4-16 h. Incorporation under these conditions was higher than 70%.

Alternatively, DNA was labelled using Stratagene Prime-It™ II random primer kit[38], with incorporation routinely being in excess of 90% of the added label.

Labelled DNA was separated from free nucleotides by chromatography through a Sephadex G-50 column[39], collecting fractions of approximately 500 μl in TE.

Probes were denatured by boiling for 5 minutes in a water bath followed by incubation for 5-10 minutes on ice immediately before adding to the hybridization solution.

2.15. Hybridization of labelled probes to DNA immobilized on nylon membranes.

Blotted nylon membranes were pre-hybridized inside a glass tube with 50-100 ml (depending on membrane size) of a

---

[38] catalogue number 300385.

[39] Nick-column, Pharmacia.

solution containing 6X SSC, 5X Denhardt's solution[40], 0.5% SDS and 100 μg/ml denatured HSD. Pre-hybridization was carried out for 2-4 h at 65°C during rotation in an incubator[41].

The pre-hybridization solution was discarded and replaced with 2-5 ml of hybridization solution[42] and incubated at 65°C for 14-16 h.

The membranes were then washed twice in 50 ml of a solution containing 2X SSC and 0.5% SDS for 15 minutes at room temperature, followed by a final wash in 50 ml of a solution containing 0.1X SSC and 0.5% SDS for 30 minutes at 65°C (repeated twice).

While still moist, the membranes were wrapped in cling film and the hybridization bands were detected by autoradiography. A sheet of film (Kodak X-Omat K or Fuji RX) was placed in contact with the membrane for a variable amount of time (from 30 minutes to 3 days) at -70°C and the film developed according to the manufacturer's recommendations.

2.16. Dephosphorylation of DNA.

For some applications vector DNA was dephosphorylated prior to ligation to the insert DNA by

---

[40] 1X Denhardt's solution contains 0.02% Ficoll™ 400000, 0.02% polyvinylpyrrolidone, 0.02% BSA (Denhardt, 1966).

[41] Bachofer.

[42] 6X SSC, 5X Denhardt's solution, 0.5% SDS, 100 μg/ml denatured HSD and 0.5 ml to 1 ml of the labelled probe.

treatment with calf intestinal phosphatase (CIP).

Dephosphorylation was performed in 10 µl reactions containing 1 mM $ZnCl_2$, 1 mM $MgCl_2$ and 10 mM Tris-Cl (pH 8.3). Variable amounts of CIP were used, ranging from 0.01 U/pmol for protruding 5' termini to 0.5 U/pmol for blunt termini. Reactions involving protruding termini were incubated at 37°C for 30 minutes. Reactions with blunt termini were incubated at 37°C for 15 minutes, another aliquot of CIP was added (containing the same amount as before) and incubation allowed to proceed for 45 minutes at 55°C. The enzyme was inactivated by adding EDTA to a final concentration of 5 mM followed by incubation at 75°C for 10 minutes and successive extractions with phenol and phenol/chloroform. Finally, DNA was precipitated with ethanol, collected by centrifugation[43] and resuspended in TE.

2.17. DNA ligation.

DNA was ligated in reactions containing 20 mM Tris-Cl (pH 7.6), 5 mM $MgCl_2$, 5 mM DTT, 50 µg/ml BSA (fraction V), 1 mM ATP and 1-3 U of T4 DNA ligase, with volumes varying from 5 to 10 µl. Variable amounts of vector and insert DNA were used, depending on the optimum ratio and concentration necessary. Reactions were incubated overnight at 16°C then used immediately for transformation or *in vitro* packaging.

---

[43] 15 krpm for 15 minutes, Heraeus Biofuge 15.

2.18. *In vitro* packaging of phage DNA.

Phage lambda DNA was packaged *in vitro* using Stratagene Gigapack II packaging extracts (Kretz *et al.*, 1989)[44] and following the protocols suggested by the manufacturer.

Essentially, the packaging extracts (prepared by ultra-sonic disruption and freezing-thawing) were thawed on ice, the ligated DNA was mixed in and the samples were incubated at room temperature for 2 hours. The titre was determined by plating serial dilutions of the packaged DNA.

2.19. Identification of phage carrying specific sequences.

Phage containing specific DNA sequences were detected by plaque lifts as described by Benton and Davis (1977).

Phage were plated in L-agar[45] at densities of up to 15000 per plate (90 mm diameter) and the plates were incubated at 37°C overnight. They were then incubated at 4°C for at least 30 minutes to harden the top-agar and circular Hybond-N membranes (82 mm diameter) were placed onto them for 30 seconds. The membranes were then transferred (DNA side up) to a Petri dish containing denaturing solution (see footnote 34) and floated on this solution for 5 minutes. Membranes were then transferred to another Petri dish containing neutralizing solution

---

[44] catalogue number 200214.

[45] L-agar contains 1% tryptone, 0.5% yeast extract, 0.5% NaCl, 1.7% agar; pH 7.2.

(see footnote 35) and floated for 5 minutes followed by a rapid wash in 2X SSC. The membranes were then air-dried for 30 minutes followed by incubation at 80°C for 2 h.

The conditions used for hybridizing probes to DNA on the membranes were identical to those previously described for Southern blotting.

Plaques corresponding to positive signals were picked with a Pasteur pipette, transferred to a tube containing 1 ml of SM buffer and incubated at 4°C overnight. Eluted phage were then plated at lower densities (less than 200 per plate) and rescreened using these same procedures. Individual, well isolated plaques were picked and eluted in 1 ml of SM buffer.

2.20. Bacterial transformation.

Transformation was performed using Stratagene Epicurian Coli XL1-Blue competent cells[46] and protocols. Briefly, competent cell were thawed on ice, the DNA was mixed and the samples heat-shocked. After incubation for 1 hour at 37°C, the cells were plated in L-agar plates containing the appropriate antibiotic (together with X-Gal and IPTG, when colour selection was necessary).

---

[46] catalogue number 200236.

2.21. Colony hybridization.

Identification of bacterial colonies carrying specific sequences was based on the technique described by Grunstein and Hogness (1975), with a few modifications.

Individual colonies were transferred simultaneously, using sterile tooth-picks, to a Hybond-N membrane placed on the surface of an L-agar plate and to a replica plate containing the appropriate antibiotic. The replica plate was incubated at 37°C overnight and then kept at 4°C for subsequent use while the plate containing the nylon membrane was incubated at 37°C for 4 h. The membrane was removed and placed (colony side up) in a Petri dish containing 10% SDS, for 3 minutes. The membrane was then subsequently transferred to Petri dishes containing denaturing and neutralizing solutions (see footnotes 34 and 35) for 5 minutes, followed by a quick wash in 2X SSC. Membranes were air-dried for 30 minutes and incubated at 80°C for 2 h in an oven.

Conditions used for hybridizing probes to DNA on the membranes were the same as previously described for Southern blotting.

2.22. Polymerase Chain Reaction (PCR).

DNA was amplified by PCR essentially as described in Innis *et al.* (1990). The reactions contained 50 mM KCl, 10 mM Tris-HCl (pH 9), 0.1% Triton X-100, 2.5 mM $MgCl_2$, 200 μM dATP, 200 μM dCTP, 200 μM dGTP, 200 μM dTTP, 0.8 μM each

primer, 10 aM to 10 fM target DNA and 2.5 U Taq polymerase, in a final volume of 25 µl. The reactions were layered with a drop of mineral oil and transferred to a thermal cycler[47].

Cycling parameters varied according to the target DNA to be amplified and can be summarized as follows:

1) initial denaturation, at 94°C for 4-5 minutes;

2) consecutive cycles of annealing (35-60°C, for 30-60 seconds), primer extension (72°C, for 1-5 minutes) and denaturation (94°C, for 30-60 seconds), repeated 30-35 times;

3) a final cycle of annealing (35-60°C, for 1-2 minutes) and primer extension (72°C, 5-10 minutes).

Whenever formation of non-specific products was a problem, the "touchdown" PCR technique was used (Don *et al.*, 1991). Basically, in this modification of the PCR technique the annealing temperature is successively reduced by 1-2°C every cycle of denaturation-annealing-extension. The higher temperatures used during the initial cycles ensure reduced spurious priming during gene amplification.

2.23. DNA sequencing.

DNA was sequenced by the chain termination method (Sanger *et al.*, 1977).

---

[47] Hybaid OmniGene.

2.23.1. Preparation of ssDNA templates.

A single bacterial colony containing the recombinant plasmid was inoculated into 5 ml L-broth (containing 12 μg/ml tetracycline and 100 μg/ml ampicillin) and incubated at 37°C overnight. 3 ml from this pre-inoculum was withdrawn, inoculated into 30 ml of superbroth[48] and incubated at 37°C for 2 h with agitation. 30 μl of helper phage (VCSM13[49], $1x10^{11}$ pfu/ml) was added and incubation allowed to proceed for 7 h at 37°C with vigorous agitation. Cultures were transferred to 65°C for 15 minutes and bacteria removed by centrifugation[50]. 12 ml from this cleared supernatant was transferred to a clean flask and phage precipitated by adding 3 ml of 3.5 M ammonium acetate, 20% PEG-8000 solution. Samples were incubated at room temperature for 15 minutes, centrifuged[51], the supernatant removed and the pellet resuspended in 3 ml TE. Samples were extracted several times with phenol/chloroform (1:1), until no interface was visible, and ssDNA precipitated by adding an equal volume of 7.5 M ammonium acetate (pH 7.5) and 2 volumes of 100% ethanol. DNA was collected by centrifugation[52], dried in a vacuum desiccator and resuspended in 10 μl TE.

---

[48] superbroth contains 3.5% Bacto-tryptone, 2% Bacto-yeast extract, 0.5% NaCl; pH 7.5.

[49] Stratagene, catalogue number 200251.

[50] 10 krpm for 5 minutes, rotor JA-14, J2-21 Beckman centrifuge.

[51] 12 krpm for 20 minutes, rotor JA-20, J2-21 Beckman centrifuge.

[52] 10 krpm for 15 minutes, rotor JA-20, J2-21 Beckman centrifuge.

2.23.2. Sequencing reactions.

ssDNA was sequenced using T7 DNA polymerase[53] (Tabor and Richardson, 1987) and α-^{35}S-dATP. Reagents and protocols were used as supplied by the manufacturer.

2.23.3. Sequencing gel electrophoresis.

Sequencing reactions were run in 6% acrylamide, 8 M urea gels[54], polymerised by the addition of 0.1 g ammonium persulphate and 50 µl TEMED. Gels were cast in 30 x 40 cm glass plates separated by a wedge spacer (0.4-0.8 mm)[55]. Prior to loading the samples, gels were pre-run at 70 W for at least 1 h. Samples were incubated at 75-80°C for 2 minutes immediately before loading. Running times varied from 2 to 8 h, depending on the fragment sizes to be resolved. After electrophoresis, gels were fixed by immersing in 10% acetic acid, 10% methanol for 1 h and transferred to the surface of a Whatman 3MM paper. Gels were then dried for 40 minutes in a gel dryer[56] and exposed overnight to Hyperfilm-ßMAX[57] at room temperature.

---

[53] Sequenase (version 2.0), United States Biochemical.

[54] 15 ml 40% acrylamide/bis-acrylamide solution (19:1), 48.05 g urea, 20 ml 10x TBE, water to 200 ml.

[55] assembled on a BRL model S2 sequencing gel kit.

[56] Bio-Rad, model 583.

[57] Amersham.

2.24. Computer analysis of nucleic acid and amino acid sequences.

Computer analysis was performed on a VAX computer using the GCG (Devereux *et al.*, 1984) or Staden (Staden, 1982, 1984abc, 1987) programmes. Oligonucleotide primers for sequencing and PCR were designed using the programme Oligo (Rychlik and Rhoads, 1989), running on a PC-compatible computer. Multiple amino acid sequence alignments were performed using the programme Clustal (Higgins and Sharp, 1988, 1989).

2.25. Production of transgenic nematodes.

Transgenic *C. elegans* was obtained by micro-injecting DNA into each arm of the hermaphrodite gonad, essentially as described by Mello *et al.* (1991).

DNA solutions for micro-injection contained 100 μg/ml pRF4 and 20 μg/ml of the transforming plasmid and were centrifuged[58] to remove any particulate matter immediately before filling the needles. Cosmids were injected at a concentration of 100 μg/ml (without the plasmid pRF4). Needles were prepared from glass capillaries[59] on a pipette puller[60] and

---

[58] 15 krpm for 5 minutes, Heraeus Biofuge 15.

[59] Clark Electromedical Instruments, item GC120TF-15.

[60] Computer Controlled Electrode Puller model 763, Campden Instruments Ltd.

loaded with DNA samples by inserting a draw-out capillary pipette down the wide bore of the needle and depositing the DNA solution just behind the tip (Fire, 1986).

Animals (large, healthy adult hermaphrodites) were immobilized by transferring them to a agarose pad[61] and covered with liquid paraffin. Usually five nematodes at a time were aligned in a row with the ventral side facing away from the needle and with enough space between worms to allow the needle access to each animal.

Slides containing the agarose pads were mounted on the stage of a microscope[62] equipped with differential interference contrast optics and the needle controlled by a micromanipulator[63]. Just prior to injection, the needle tips were broken open by gently touching dust particles embedded in the agar pad. Immediately after puncturing the worm's cuticle and positioning the needle inside the gonad, nitrogen pressure (2-3 atmospheres) was applied and maintained until the entire gonad was filled with DNA solution. After the last animal on the agar pad was injected, a drop of sterile M9 buffer was applied and the worms transferred to a seeded NGM plate and incubated at 20°C.

In those injections including the plasmid pRF4, the F1

---

[61] prepared by depositing a drop of 3% agarose on a glass slide, immediately covering with a coverslip and drying at 80°C for 30 minutes (Fire, 1986).

[62] Olympus IMT-2, mounted on a Wentworth Labs Ltd vibration isolation table, model AVT-201.

[63] Narishige joystick hydraulic micromanipulator, model MO-202.

transformants exhibiting roller phenotype, in groups of five, were transferred to seeded plates. The F2 transformants were then individually transferred to seeded plates and F3 progeny originating from the animals giving rise to the largest number of roller worms were maintained. Animals injected with cosmids (without pRF4) were simply observed for the rescue of the dumpy phenotype in the progeny.

2.26. Detection of chimeric ß-galactosidase expression in transgenic nematodes.

The histochemical staining procedure used was a modification of the technique described by Fire (1990) and further modified by Hope (1991).

Worms were transferred to a drop of distilled water on a glass slide and gently covered with a coverslip. Slides were then placed on the surface of a metal block half immersed in liquid Nitrogen, for 30 seconds. The coverslip was flipped off with a scalpel blade and the slides were placed in 100% methanol, kept at -20°C, for 5 minutes. Slides were then transferred to 100% acetone, also at -20°C, for 5 minutes and the preparation removed and allowed to air dry at room temperature for 5-10 minutes. When completely dried, 100 µl staining solution[64] was applied, the preparations were covered with a coverslip (avoiding air bubbles) and sealed to the slide with nail varnish. The slides

---

[64] staining solution contains 0.2 M $Na4P_2O7$, 1 µM $MgCl_2$, 5 mM $K4Fe(CN)_6$, 5 mM $K3Fe(CN)_6$, 0.004% SDS, 0.032% X-Gal.

were incubated at 37°C for a period ranging from 4 hours to overnight before observation on a microscope.

2.27. Preparation of *C. elegans* cuticle extracts.

Worms were grown in a 90 mm NGM plate until all the bacteria had been consumed (approximately 1 week) and harvested by washing the plate twice with M9. The worms were then centrifuged[65], the supernatant was discarded and the pellet was resuspended in 5 ml Calcium buffer[66]. Finally, this last step was repeated and the worm-containing pellet was stored at -70°C.

Cuticles were isolated by a modification of the method described by Cox *et al.* (1981c). Frozen worms were thawed on ice and disrupted by sonication using 15 pulses of 20 seconds (each followed by a 20 seconds interval), in a sonicator. During all manipulations the cuticles were kept on ice. The cuticles were centrifuged[67], then the supernatant was removed and concentrated by spin-dialysis in a Centricon-10 microconcentrator to a final volume of 40 µl.

Two different methods were used to extract the collagenolytic activity associated with the cuticle material. The first method relied on the dissociation of the activity from the

---

[65] 3 krpm for 5 minutes (4°C), rotor TH-4, TJ-6 Beckman centrifuge.

[66] Calcium buffer contains 10 mM Tris-Cl (pH 7.5), 5 mM $CaCl_2$.

[67] 15 krpm for 2 minutes at 4°C, Heraeus Biofuge 15.

cuticle collagen by treatment with a high Calcium concentration (Harris and Vater, 1982). The pellet containing the cuticles was resuspended in 100 µl extraction buffer[68] and incubated on ice for 2 h. The sample was centrifuged[69] and the supernatant used for the enzymatic assays. In the second method, the cuticle pellet was resuspended in 100 µl 5% ß-ME and incubated for 15 minutes at room temperature, followed by centrifugation as described above for the Calcium extracts.

2.28. Substrate gel electrophoresis.

Collagenolytic activity in cuticle extracts of *C. elegans* was assayed using a modification of the method described by Heussen and Dowdle (1980), with collagen being copolymerised in a standard SDS-PAGE (Laemmli, 1970).

Polyacrylamide gels were polymerised in the presence of collagen[70]. Boiling and reducing agents were not used, otherwise the procedure was that of Laemmli (1970), with a stacking gel not containing any substrate[71]. Gels were prepared

---

[68] extraction buffer contains 10 mM Tris-Cl pH 7.5, 0.1 M $CaCl_2$, 2.5% Triton-X-100.

[69] 15 krpm for 15 minutes at 4°C, Heraeus Biofuge 15.

[70] 7.5% acrylamide/bis-acrylamide (29:1), 0.1% SDS, 0.375 M Tris-Cl pH 8.8, 0.1% collagen (acid soluble, type III, Sigma). The gel was polymerised by adding 10 mg ammonium persulphate and 5 µl TEMED.

[71] 4% acrylamide/bis-acrylamide (29:1), 0.1% SDS, 0.125 M Tris-Cl pH 6.8, The gel was polymerised by adding 5 mg ammonium persulphate and 5 µl TEMED.

on a Bio-Rad[72] kit (0.75 mm spacer thickness). A Tris-glycine running buffer[73] was used. Samples from the cuticle extracts (20 µl) were diluted 1:1 in a sample buffer[74] and loaded in the wells. Electrophoresis was carried out at 4°C, 100 V for 30 minutes in the stacking gel and 200 V for 50 minutes in the resolving gel.

After electrophoresis, the gel was washed twice for 30 minutes in assay buffer[75] and incubated overnight at 37°C in this same buffer. The gels were then stained with Coomassie Brilliant Blue R250 for 3-4h and destained in 45% methanol with 10% acetic acid until bands of collagen degradation were visualized.

Substrate gels incorporating casein were also used to assay for substrate specificity. The protocols used were the same as described before for gels containing collagen, with casein (0.1% final concentration) instead of collagen being copolymerised in the gels.

---

[72] Mini-PROTEAN™ II dual slab cell.

[73] running buffer contains 15 g/l Tris-OH, 72 g/l glycine, 5 g/l SDS, pH 8.3.

[74] sample buffer contains 20% glycerin, 0.025% bromophenol blue, 125 mM Tris-Cl (pH 6.8), 0.4% HEPES.

[75] assay buffer contains 50 mM Tris-Cl pH 7.4, 5 mM $CaCl_2$, 2.5% Triton-X-100.

3. RESULTS.

3.1. Identification of a collagenolytic activity in *C. elegans* and attempts at cloning the corresponding gene.

3.1.1. Detection of a collagenolytic activity in cuticle extracts of *C. elegans*.

Cuticle extracts, as well as the supernatant from sonicated cuticles, were assayed for collagenolytic activity using substrate gels. A similar procedure, or modifications of it, has been used successfully to identify proteases in organisms like *Onchocerca* (Lackey *et al.*, 1989), *Strongyloides stercoralis* (McKerrow *et al.*, 1990) and *Schistosoma* (McKerrow *et al.*, 1985). By using this method both the number of collagenases and their nonreduced molecular weight can be determined simultaneously.

Figure 1A shows a typical result obtained when applying this assay to cuticle extracts of *C. elegans*. A band of collagen degradation can be detected in the Calcium extracts (lane 2). Although the migration behaviour of the proteins is not necessarily proportional to their molecular weights, the band of collagenolytic activity from *C. elegans* seems to comigrate with one of the bacterial activities. This activity is exclusively associated with the cuticle material and is not found in the concentrated supernatant. The *C. elegans* collagenase activity, estimated from the amount of collagen degradation, is

comparable to the bacterial collagenase, which is $1x10^{-5}$ Units[1], suggesting that it is present only in a very small quantity. This would prevent attempts to purify sufficient enzyme to determine amino acid sequence.

No substrate degradation is observed when Calcium extracts are assayed using casein gels (Figure 1B), indicating that the enzyme does not have a non-specific protease activity.

3.1.2. Designing of oligonucleotides for detecting specific sequences in metalloproteases.

As it would not be possible to obtain sufficient protein for determining the amino acid sequence of the putative *C. elegans* collagenase, a different approach was devised. It involved analysing the sequence of several metalloproteases in order to identify conserved sequences.

Initially, a search in the GenBank and EMBL databases was performed and several sequences representative of metalloproteases from different organisms were selected. These sequences were aligned using the programme Clustal (Figure 2). From the comparison of these sequences one can conclude that the overall homology among the mammalian sequences is high, and that several sites within the sea urchin collagenase are also conserved, despite the evolutionary distance. Particularly important are two sites within these

---

[1] 1 Unit liberates peptides from native collagen (type I, Sigma) equivalent in ninhydrin colour to 1 μmole of L-leucine in 5 h at pH 7.4 at 37°C in the presence of Calcium ions.

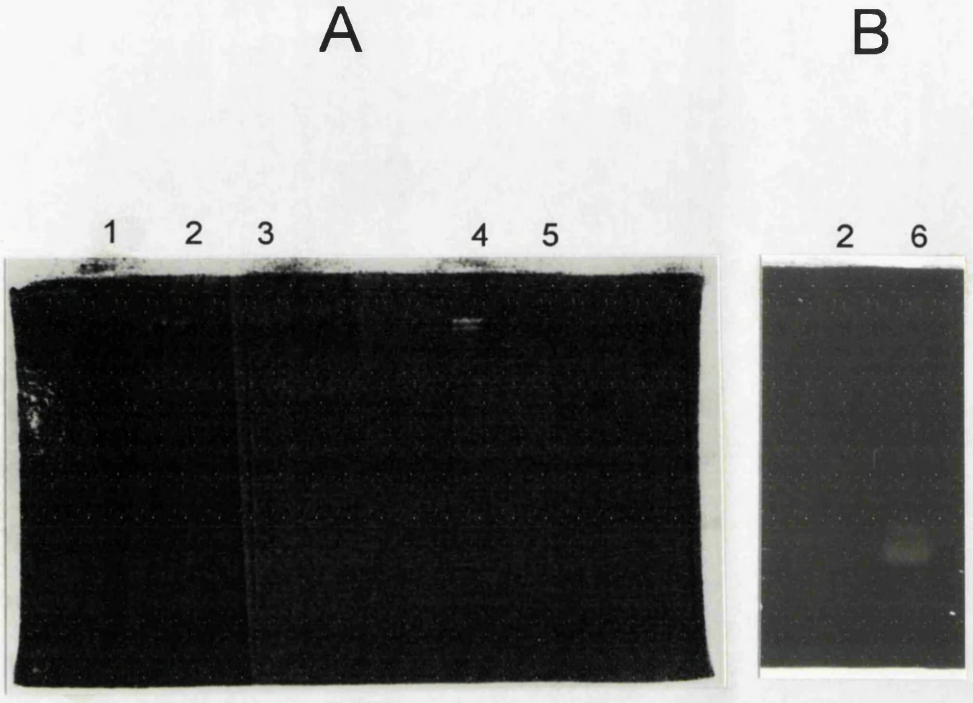
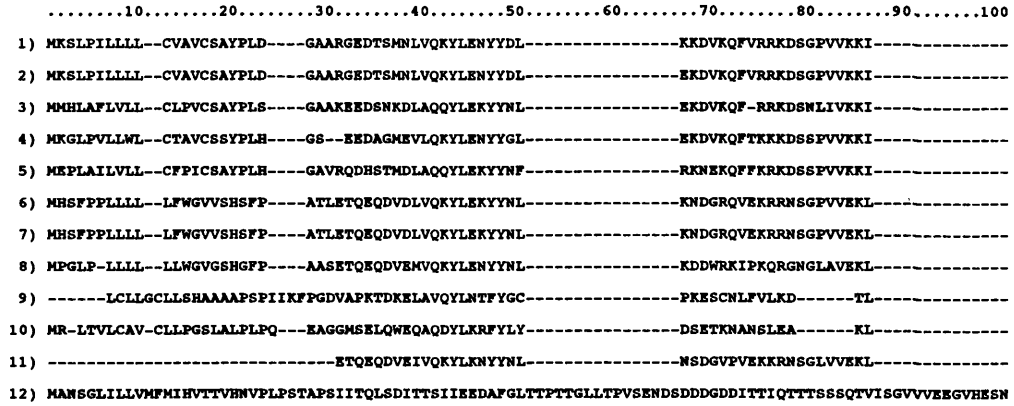
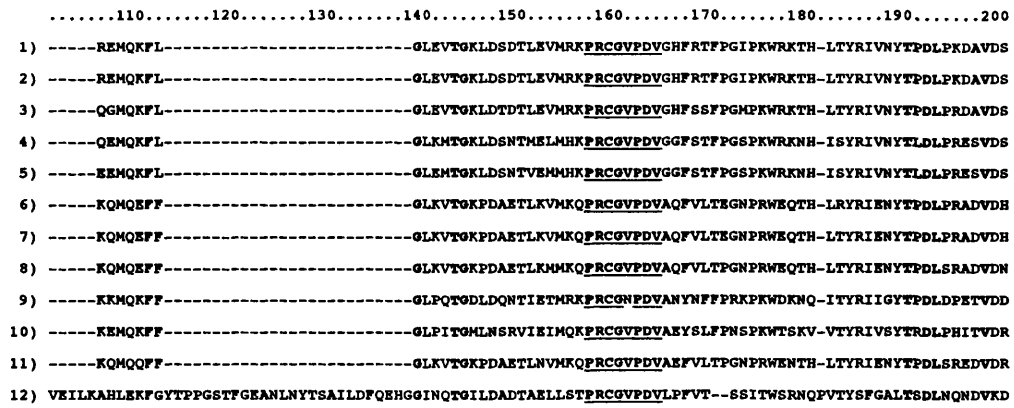


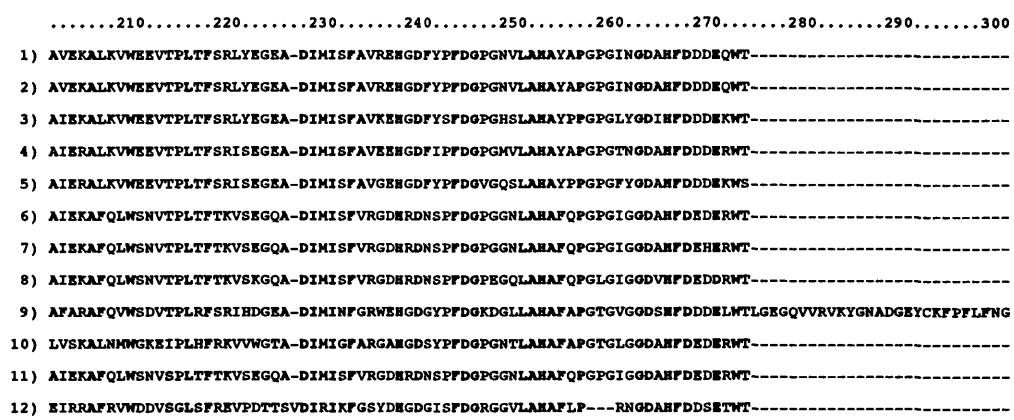
Figure 1. Detection of a collagenolytic activity in cuticle extracts of *C. elegans*.

Samples were electrophoresed in substrate gels containing collagen (panel A) or casein (panel B) which were then incubated in assay buffer (see Materials and Methods) and stained with Coomassie Blue.

1) concentrated supernatant; 2) Calcium extract from cuticles; 3) ß-ME extract from cuticles; 4) 20 µl *Clostridium histolyticum* collagenase (Sigma, type VII), $1x10^{-5}$ U (5 pg); 5) 20 µl *Clostridium histolyticum* collagenase, $1x10^{-6}$ U (0.5 pg); 6) proteinase K (Sigma, product number P2308), $1x10^{-5}$ U (1 ng).


.....10.....20.....30.....40.....50.....60.....70.....80.....90.....100
1) MKSLPILLLL--CVAVCSAYPLD----GAARGEDTSMNLVQKYLENYYDL----------KKDVKQFVRRKDSGPVVKKI----------
2) MKSLPILLLL--CVAVCSAYPLD----GAARGEDTSMNLVQKYLENYYDL----------EKDVKQFVRRKDSGPVVKKI----------
3) MMHLAFLVLL--CLPVCSAYPLS----GAAKEEDSNKDLAQQYLEKYYNL----------EKDVKQF-RRKDSNLIVKKI----------
4) MKGLPVLLWL--CTAVCSSYPLH----GS--EEDAGMEVLQKYLENYYGL----------EKDVKQFTKKKDSSPVVKKI----------
5) MEPLAILVLL--CFPICSAYPLH----GAVRQDHSTMDLAQQYLEKYYNF----------RKNEKQFFKRKDSSPVVKKI----------
6) MHSFPPLLLL--LFWGVVSHSFP----ATLETQEQDVDLVQKYLEKYYNL----------KNDGRQVEKRRNSGPVVEKL----------
7) MHSFPPLLLL--LFWGVVSHSFP----ATLETQEQDVDLVQKYLEKYYNL----------KNDGRQVEKRRNSGPVVEKL----------
8) MPGLP-LLLL--LLWGVGSHGFP----AASETQEQDVEMVQKYLENYYNL----------KDDWRKIPKQRGNGLAVEKL----------
9) ------LCLLGCLLSHAAAAPSPIIKFPGDVAPKTDKELAVQYLNTFYGC----------PKESCNLFVLKD------TL----------
10) MR-LTVLCAV-CLLPGSLALPLPQ---EAGGMSELQWEQAQDYLKRFYLY----------DSETKNANSLEA------KL----------
11) ----------ETQEQDVEIVQKYLKNYYNL----------NSDGVPVEKKRNSGLVVEKL----------
12) MANSGLILLVMFMIHVTTVHNVPLPSTAPSIITQLSDITTSIIEEDAFGLTTPTTGLLTPVSENDSDDDGDDITTIQTTTSSSQTVISGVVVEEGVHESN


.....110.....120.....130.....140.....150.....160.....170.....180.....190.....200
1) -----REMQKFL----------GLEVTGKLDSDTLEVMRKPRCGVPDVGHFRTFPGIPKWRKTH-LTYRIVNYTPDLPKDAVDS
2) -----REMQKFL----------GLEVTGKLDSDTLEVMRKPRCGVPDVGHFRTFPGIPKWRKTH-LTYRIVNYTPDLPKDAVDS
3) -----QGMQKFL----------GLEVTGKLDTDTLEVMRKPRCGVPDVGHFSSFPGMPKWRKTH-LTYRIVNYTPDLPRDAVDS
4) -----QEMQKFL----------GLKMTGKLDSNTMELMHKPRCGVPDVGGFSTFPGSPKWRKNH-ISYRIVNYTLDLPRESVDS
5) -----EEMQKFL----------GLEMTGKLDSNTVEMMHKPRCGVPDVGGFSTFPGSPKWRKNH-ISYRIVNYTLDLPRESVDS
6) -----KQMQEFF----------GLKVTGKPDAETLKVMKQPRCGVPDVAQFVLTEGNPRWEQTH-LRYRIENYTPDLPRADVDH
7) -----KQMQEFF----------GLKVTGKPDAETLKVMKQPRCGVPDVAQFVLTEGNPRWEQTH-LTYRIENYTPDLPRADVDH
8) -----KQMQEFF----------GLKVTGKPDAETLKMMKQPRCGVPDVAQFVLTPGNPRWEQTH-LTYRIENYTPDLSRADVDN
9) -----KKMQKFF----------GLPQTGDLDQNTIETMRKPRCGNPDVANYNFFPRKPKWDKNQ-ITYRIIGYTPDLDPETVDD
10) -----KEMQKFF----------GLPITGMLNSRVIEIMQKPRCGVPDVAEYSLFPNSPKWTSKV-VTYRIVSYTRDLPHITVDR
11) -----KQMQQFF----------GLKVTGKPDAETLNVMKQPRCGVPDVAEFVLTPGNPRWENTH-LTYRIENYTPDLSREDVDR
12) VEILKAHLEKFGYTPPGSTFGEANLNYTSAILDFQEHGGINQTGILDADTAELLSTPRCGVPDVLPFVT--SSITWSRNQPVTYSFGALTSDLNQNDVKD


.....210.....220.....230.....240.....250.....260.....270.....280.....290.....300
1) AVEKALKVWEEVTPLTFSRLYEGEA-DIMISFAVREHGDFYPFDGPGNVLAHAYAPGPGINGDAHFDDDEQWT----------
2) AVEKALKVWEEVTPLTFSRLYEGEA-DIMISFAVREHGDFYPFDGPGNVLAHAYAPGPGINGDAHFDDDEQWT----------
3) AIEKALKVWEEVTPLTFSRLYEGEA-DIMISFAVKEHGDFYSFDGPGHSLAHAYPPGPGLYGDIHFDDDEKWT----------
4) AIERALKVWEEVTPLTFSRISEGEA-DIMISFAVEEHGDFIPFDGPGMVLAHAYAPGPGTNGDAHFDDDERWT----------
5) AIERALKVWEEVTPLTFSRISEGEA-DIMISFAVGEHGDFYPFDGVGQSLAHAYPPGPGFYGDAHFDDDEKWS----------
6) AIEKAFQLWSNVTPLTFTKVSEGQA-DIMISFVRGDHRDNSPFDGPGGNLAHAFQPGPGIGGDAHFDEDERWT----------
7) AIEKAFQLWSNVTPLTFTKVSEGQA-DIMISFVRGDHRDNSPFDGPGGNLAHAFQPGPGIGGDAHFDEHERWT----------
8) AIEKAFQLWSNVTPLTFTKVSKGQA-DIMISFVRGDHRDNSPFDGPEGQLAHAFQPGLGIGGDVHFDEDDRWT----------
9) AFARAFQVWSDVTPLRFSRIHDGEA-DIMINFGRWEHGDGYPFDGKDGLLAHAFAPGTGVGGDSHFDDDELWTLGEGQVVRVKYGNADGEYCKFPFLFNG
10) LVSKALNMWGKEIPLHFRKVVWGTA-DIMIGFARGAHGDSYPFDGPGNTLAHAFAPGTGLGGDAHFDEDERWT----------
11) AIEKAFQLWSNVSPLTFTKVSEGQA-DIMISFVRGDHRDNSPFDGPGGNLAHAFQPGPGIGGDAHFDEDERWT----------
12) EIRRAFRVWDDVSGLSFREVPDTTSVDIRIKFGSYDHGDGISFDGRGGVLAHAFLP---RNGDAHFDDSETWT----------

```
    .......310.......320.......330.......340.......350.......360.......370.......380.......390.......400
 1) ----------------------------------------------------------------------------KDTT----------------------------
 2) ----------------------------------------------------------------------------KDTT----------------------------
 3) ----------------------------------------------------------------------------EDAS----------------------------
 4) ----------------------------------------------------------------------------DDVT----------------------------
 5) ----------------------------------------------------------------------------LGPS----------------------------
 6) ----------------------------------------------------------------------------NNFR----------------------------
 7) ----------------------------------------------------------------------------NNFT----------------------------
 8) ----------------------------------------------------------------------------KDFR----------------------------
 9) KEYNSCTDTGRSDGFLWCSTTYNFEKDGKYGFCPHEALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPETAMSTVG
10) ----------------------------------------------------------------------------DGSS----------------------------
11) ----------------------------------------------------------------------------KNFR----------------------------
12) ----------------------------------------------------------------------------EGTR----------------------------

    .......410.......420.......430.......440.......450.......460.......470.......480.......490.......500
 1) ------------------------------------------------------GTNLFLVAA-HEIGHSLGLFHSANTEALMYPLYHSLTDLTRFRLSQD
 2) ------------------------------------------------------GTNLFLVAA-HEIGHSLGLFHSANTEALMYPLYHSLTDLTRFRLSQD
 3) ------------------------------------------------------GTNLFLVAA-HELGHSLGLFHSANTEALMYPLYNSFTELAQFRLSQD
 4) ------------------------------------------------------GTNLFLVAA-HELGHSLGLFHSANAEALMYPVYKSSTDLARFHLSQD
 5) ------------------------------------------------------GTNLFLVAA-HELGHSLGLFHSNNKESLMYPVYRFSTSQANIRLSQD
 6) ------------------------------------------------------EYNLHRVAA-HELGHSLGLSHSTDIGALMYPSYTFSGDV---QLAQD
 7) ------------------------------------------------------EYNLHRVAA-HELGHSLGLSHSTDIGALMYPSYTFSGDV---QLAQD
 8) ------------------------------------------------------NYNLYRVAA-HELGHSLGLSHSTDIGALMYPNYMFSGDV---QLAQD
 9) GNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATTANYDDDRKWGFCPDQ-GYSLFLVAA-HEFGHAMGLEHSQDPGALMAPIYT---YTKNFRLSQD
10) ------------------------------------------------------LGIN-FLYAATHELGHSLGMGHSSDPNAVMYPTYG-NGDPQNFKLSQD
11) ------------------------------------------------------DYNLYRVAA-HELGHSLGLSHSTDIGALMYPNYIYTGDV---QLSQD
12) ------------------------------------------------------SGTNLFQVAA-HEFGHSLGLYHSTVRSALMYPYYQ--GYVPNFRLDND

    .......510.......520.......530.......540.......550.......560.......570.......580.......590.......600
 1) DINGIQSLYGPPPD------SPETPLVP------------------------TEPVPPEPGTPANCDPALSFDAVSTLRGEILI-FKDRHFWRK-SLRK
 2) DINGIQSLYGPPPD------SPETPLVP------------------------TEPVPPEPGTPANCDPALSFDAVSTLRGEILI-FKDRHFWRK-SLRK
 3) DVNGIQSLYGPPPA------STEEPLVP------------------------TKSVPSGSEMPAKCDPALSFDAISTLRGEYLF-FKDRYFWRR-SHWN
 4) DVDGIQSLYGPPTE------SPDVLVVP------------------------TKSNSLDPETLPMCSSALSFDAVSTLRGEVLF-FKDRHFWRK-SLRT
 5) DIEGIQSLYGARP-------SSDATVVP------------------------VPSVSPKPETPVKCDPALSFDAVTMLRGEFLF-FKDRHFWRR-TQWN
 6) DIDGIQAIYGRSQN------PVQ--------------------------------PIGPQTPKACDSKLTFDAITTIRGEVMF-FKDRFYMRT-NPFY
 7) DIDGIQAIYGRSQN------PVQ--------------------------------PIGPQTPKACDSKLTFDAITTIRGEVMF-FKDRFYMRT-NPFY
 8) DIDGIQAIYGPSQN------PSQ--------------------------------PVGPQTPKVCDSKLTFDAITTIRGEIMF-FKDRFYMRA-NPYY
 9) DIKGIQELYGASPDIDLGTGPTP-----------------------------T----LGPVTPEICKQDIVFDGIAQIRGEIFF-FKDRFIWRTVTPRD
10) DIKGIQKLYGKR--------------------------------------------------------------------------------------
11) DIDGIQAIYGPSEN------PVQ--------------------------------PSGPQTPQVCDSKLTFDAITTLRGELMF-FKDRFYMRT-NSFY
12) DIAGIRSLYGSNSG------SGTTTTTRRPTTTRATTTRRTTTTRATTTRATTTTTTSPSRPSPPRRACSGSFDAVVRDSSNRIYALTGPYFWQL-DQ--
```

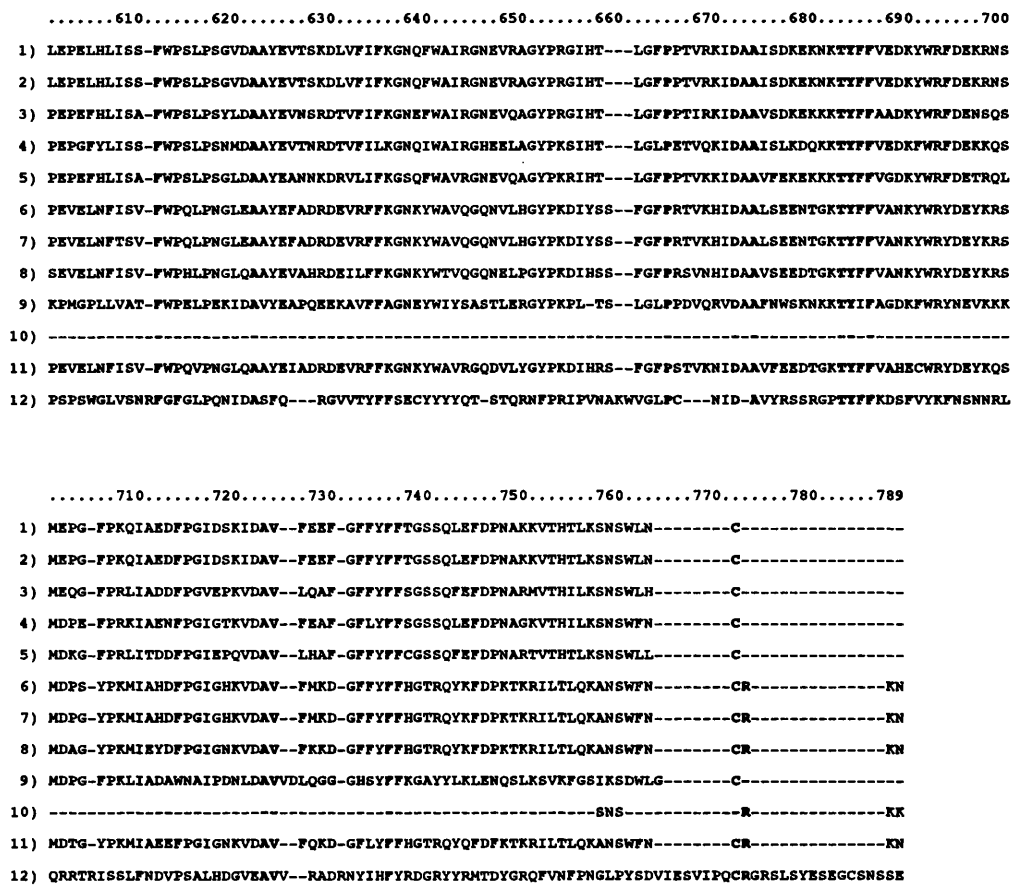


Figure 2. Amino acid sequence alignment of metalloproteases from different organisms.

The sequences were aligned using the programme Clustal (Higgins and Sharp, 1988, 1989). Gaps, represented by (-), were introduced to maximize homology. The two most conserved sequences, the activation site (position 157) and the active centre of enzyme (position 460) are underlined.

1) human stromelysin (Whitham *et al.*, 1986); 2) human metalloprotease-3 (Saus *et al.*, 1988); 3) human stromelysin-2 (Muller *et al.*, 1988); 4) rat transin (Matrisian *et al.*, 1985); 5) rat transin-2 (Breathnach *et al.*, 1987); 6) human skin collagenase (Goldberg *et al.*, 1986); 7) human collagenase (Whitham *et al.*, 1986); 8) rabbit collagenase (Fini *et al.*, 1987); 9) human type-IV collagenase (Collier *et al.*, 1988); 10) human pump-1 metalloprotease (Muller *et al.*, 1988); 11) porcine type I collagenase (Krebs *et al.*, 1990); 12) *Paracentrotus lividus* (sea urchin) collagenase (Lepage and Gache, 1990).

sequences that are very well conserved (highlighted in Figure 2).

The first of these sites, near the protein amino-terminal end, has the consensus sequence "PRCGVPDV", which is completely conserved in all but one of the sequences analysed. It has been proposed that this sequence, called the activation site, is involved in regulating the enzyme activity. The second site is the active centre of the enzyme. It is closer to the carboxy-terminal end and contains the Zinc-binding region. The consensus sequence for this site is: "VAAHE(I/L/F)GH(S/A)(L/M)G(L/M)(F/S/G/Y/E/)HS".

Although there are no crystallographic data available, it has been proposed that two of the histidines in the active centre act together in coordinating the Zinc atom, essential for the enzyme activity. According to a recent model, in a newly secreted metalloprotease the cysteine present in the activation site interacts with the Zinc atom in the active centre, thus rendering inactive the enzyme (Van Wart and Birkedal-Hansen, 1990; Matrisian, 1990, 1992). It would be only after the removal of the amino-terminal part, including the activation site, that the enzyme would become active.

Based on the amino acid sequence of these conserved regions two oligonucleotides, MSO-S and MSO-A, were designed (Figure 3) to be used in PCR and hybridization experiments. These oligonucleotides were designed incorporating inosine in some positions in order to avoid synthesizing highly degenerated sequences. Inosine forms stable base pairs with all four conventional bases, and the strength of the pairing is approximately equal in each case (Martin *et al.*, 1986; Ohtsuka *et al.*, 1985; Takahashi *et al.*, 1985).

MSO-S oligonucleotide

```
                                        N
                P     R     C     G     V     P     D     V
5' AAGGAT CCI CGI TGT GGI GTI CCI GAT GT   3'
     BamHI    A     C     AA          C
```

MSO-A oligonucleotide (complementary sequence)

```
                                        I
                                        F
          V     A     A     H     E     L     G     H
5' GTN GCN GCI CAT GAA ITI GGI CA   3'
                    C   G
```

MSO-A oligonucleotide (actual sequence)

```
5' AAAAGCT TG ICC IAI TTC ATG IGC AGC AAC 3'
     HindIII          C   G       C   C
                                  G   G
                                  T   T
```

Figure 3. Oligonucleotides specific for conserved sequences within metalloprotease sequences.

The oligonucleotides are presented below the corresponding amino acid sequences (one letter code) from the selected regions in the metalloproteases (starting at positions 157 and 460, respectively, in Figure 2). The sequence for oligonucleotide MSO-A was reversed respectively to its amino acid sequence so that it would hybridize to the anti-sense DNA strand.

3.1.3. Attempts to isolate the putative *C. elegans* collagenase gene by PCR.

Two different sets of PCR reactions were performed using the MSO-S and MSO-A oligonucleotides as primers for amplifying specific sequences in human, sea urchin and *C. elegans* DNA[2]. *E. coli* OP50 DNA was also included as a negative control. In the first set of reactions (Figure 4A), 5 cycles of annealing-extension-denaturation were performed using an annealing temperature of 37°C followed by 30 cycles at 50°C. In the second set (Figure 4B), 2 cycles were done at 42°C followed by 35 cycles at 55°C. No specific PCR product amplified from *C. elegans* DNA could be identified using these conditions in the PCR reactions.

In order to assess the efficiency of the oligonucleotides being used as well as the target sequences, two additional control experiments were performed. Each control consisted of three PCR reactions. The first reaction contained 6p6.1 DNA (cloned sea urchin collagenase). The second reaction contained sea urchin or *C. elegans* genomic DNA, and the third contained a mixture of 6p6.1 DNA and sea urchin or *C. elegans* DNA. DNA was amplified using 5 cycles of amplification with an annealing temperature of 37°C followed by 30 cycles at 50°C. Aliquots were withdrawn and run in an agarose gel (Figure 5). Under the conditions used no specific product was amplified when using either sea urchin or *C. elegans* DNA. However, a PCR product of

---

[2] the human (HeLa cells) and sea urchin DNAs were a gift from H. Jacobs (Genetics Department, Glasgow University).

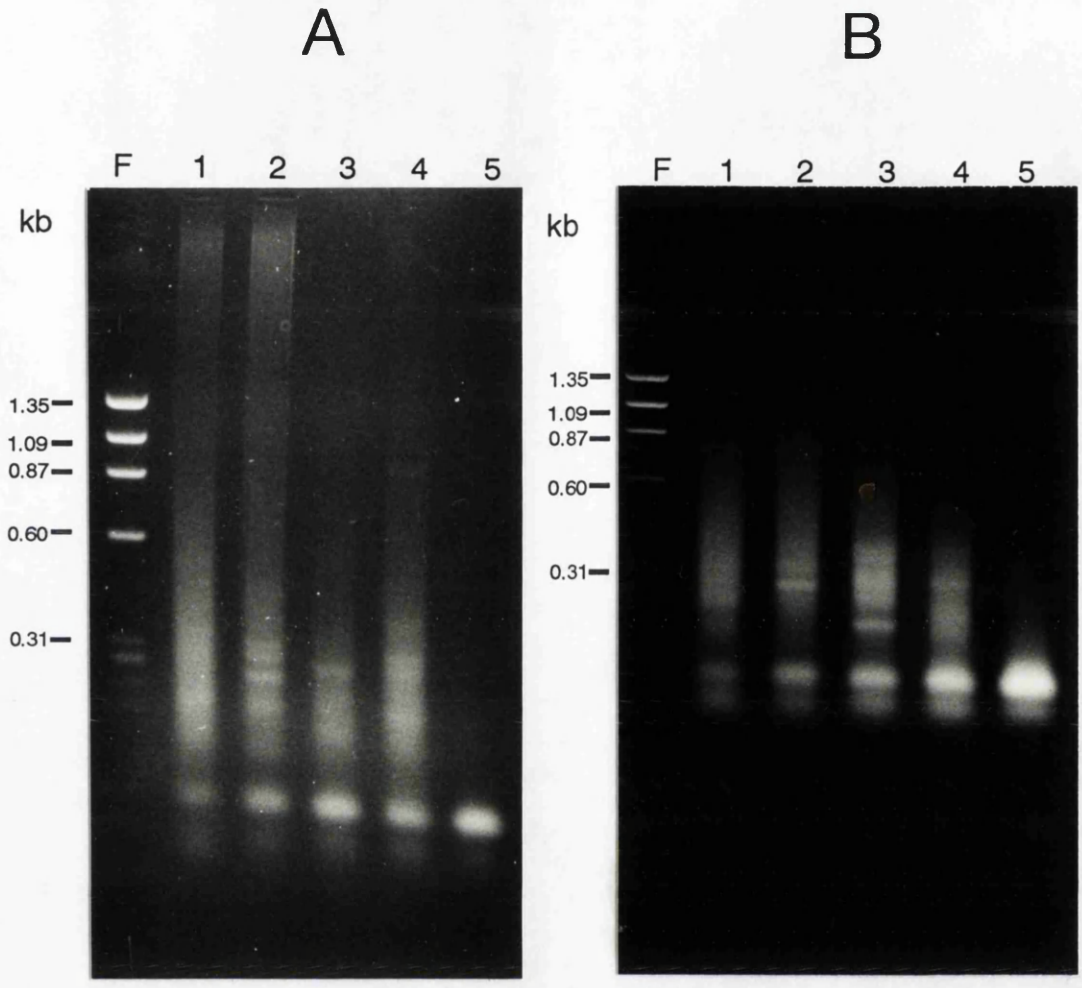


Figure 4. Agarose gel electrophoresis of PCR products amplified using MSO-S and MSO-A primers.

Conditions used in all PCR reactions were: 30 seconds denaturation (94°C) followed by 30 seconds annealing and 1 minute extension (72°C). PCR was performed using $5x10^{-7}$ pmol of each target DNA. In panel A, 5 cycles of annealing-extension-denaturation were performed using an annealing temperature of 37°C followed by 30 cycles at 50°C. In panel B, 2 cycles were performed at 42°C followed by 35 cycles at 55°C. 6 µl samples (corresponding to 1/4 of the PCR reactions) were electrophoresed in a 1.8% agarose gel.
F) ΦX-174/*Hae*III; 1) human DNA template (1 µg); 2) sea urchin DNA (270 ng); 3) *C. elegans* DNA (70 ng); 4) *E. coli* (OP50 strain) DNA (1.3 ng); 5) no DNA added to PCR.

expected length (approximately 400 bp) could be detected when using the cloned sea urchin collagenase as target sequence, indicating that the primers were used efficiently. Moreover, this same product could be identified in the PCRs where the DNA from the cloned sea urchin collagenase and genomic DNAs were combined, suggesting that a non-specific inhibition could not account for the negative results when using only genomic DNA.

Since it was not possible to amplify any metalloprotease encoding sequence from *C. elegans* genomic DNA, another approach was devised using DNA from a *C. elegans* cDNA library, constructed in λZAP, as target for PCR amplification. To isolate DNA, samples corresponding to $1x10^7$ and $1x10^4$ pfu were withdrawn from the library and heated at 80°C for 10 minutes followed by freezing at -70°C for 10 minutes. This heat/freeze cycle was repeated once before the samples were used in the PCR reactions. The reactions consisted of five different sets, each set using a different combination of primers. In addition to using primers specific for the two conserved sites in the metalloprotease sequences (MSO-S and MSO-A), the T3[3] and T7[4] primers were also used. This strategy would enable the amplification of a product if the target sequence for one of the MSO primers was either absent or modified. The first set of PCRs contained primers MSO-S and MSO-A, set two contained primers MSO-A and T3, set three contained primers MSO-S and

---

[3] this primer hybridizes to the T3 promoter flanking the polylinker of λZAP and has the sequence 5' ATTAACCCTCACTAAAG 3'.

[4] this primer hybridizes to the T7 promoter flanking the polylinker of λZAP and has the sequence 5' AATACGACTCACTATAG 3'.

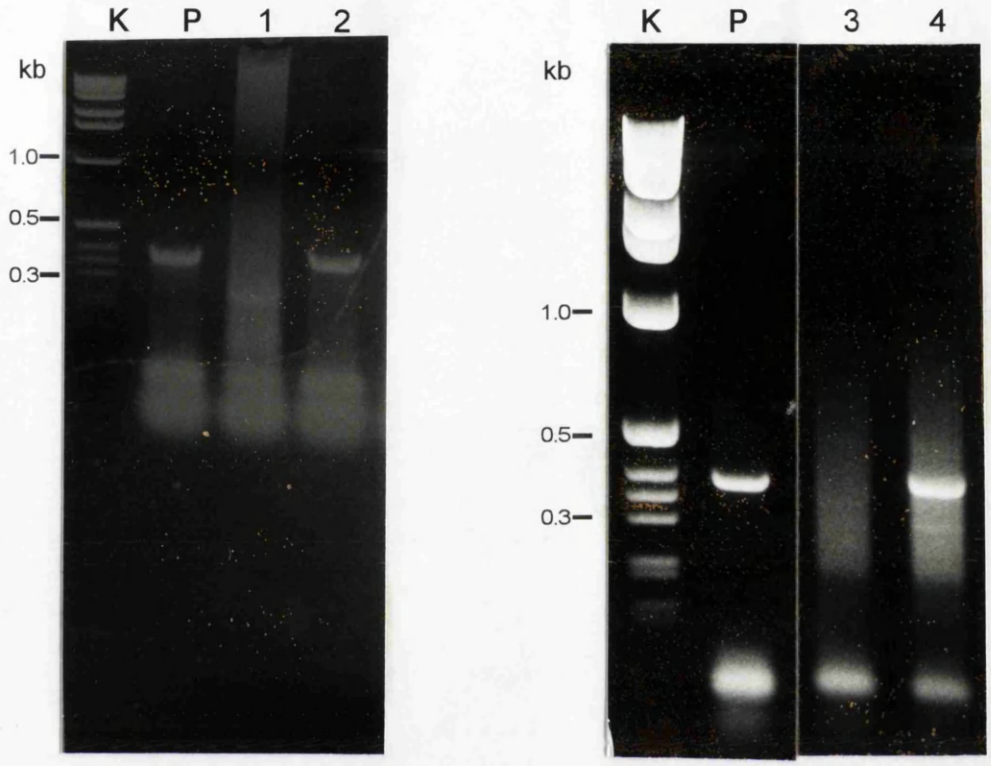


Figure 5. Agarose gel electrophoresis of PCR products amplified using sea urchin, *C. elegans* and 6p6.1 plasmid DNA.

Conditions used in all PCR reactions were: 30 seconds denaturation (94°C) followed by 30 seconds annealing and 1 minute extension (72°C). 5 cycles of annealing-extension-denaturation were done using an annealing temperature of 37°C followed by 30 cycles at 50°C. PCR was performed using $5x10^{-7}$ pmol of each target DNA. 10 µl samples from the PCR reactions were electrophoresed in a 1.8% agarose gel.

K) 1 kb ladder; P) 6p6.1 DNA template (1.77 pg); 1) sea urchin DNA (270 ng); 2) sea urchin plus 6p6.1 DNA; 3) *C. elegans* DNA (70 ng); 4) *C. elegans* plus 6p6.1 DNA.

T3, set four contained primers MSO-A and T7, and set five contained primers MSO-S and T7.

Each of the five sets comprised three separate reactions, using the equivalent of $1x10^7$ pfu of phage, $1x10^4$ pfu of phage, and a control with no phage DNA added. The "touchdown" PCR technique was used, with annealing temperatures varying from 60°C to 45°C over 16 cycles (followed by 20 cycles at 45°C).

After the PCRs were completed, 10 µl samples were electrophoresed in a 1.5% agarose gel (Figure 6A). A PCR product, with a size just below 1 kb, was detected in the reaction using $1x10^7$ pfu of phage and oligonucleotides MSO-S and MSO-A as primers (Figure 6A, reaction 1W). In order to verify if this PCR product was being amplified from cloned sequences in λZAP or from sequences in the vector, the PCR was repeated using the same conditions described before and including a control reaction containing 5 ng λZAP as target DNA[5]. As shown in Figure 6B, all PCR products were amplified exclusively from λZAP sequences.

3.1.4. Hybridization of MSO-S and MSO-A oligonucleotides to *C. elegans* genomic DNA.

Another approach used to try to identify the putative *C. elegans* metalloprotease encoding sequence was by hybridization of labelled MSO-S and MSO-A oligonucleotides to

---

[5] $1x10^8$ molecules of phage lambda DNA have a mass of 4.5 ng.

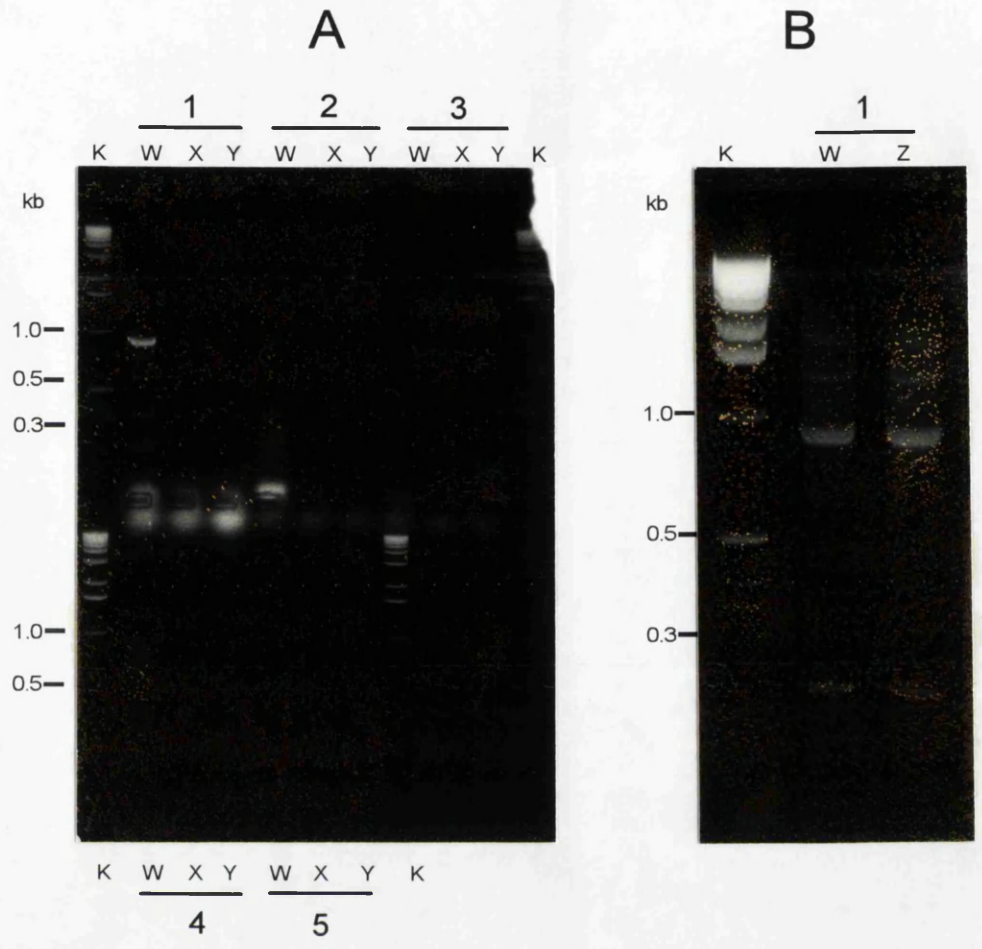


Figure 6. Agarose gel electrophoresis of PCR products amplified using DNA from a λZAP/*C. elegans* cDNA library.

"Touchdown" PCR was performed using 1 minute denaturation (94°C), 1 minute annealing (with temperature varying from 60°C to 45°C over 16 cycles, followed by 20 cycles at 45°C) and 2 minutes extension (72°C). Each of the PCR reactions contained $1x10^7$ pfu (W), $1x10^4$ pfu (X), and no DNA (Y), respectively (panel A). The control PCR in panel B (lane Z) contained 5 ng of λZAP DNA and was performed using the same conditions as described for panel A. 10 µl samples were withdrawn from the PCRs and electrophoresed in a 1.5% agarose gel. Each set of PCRs was performed using the following primers: 1) MSO-S and MSO-A; 2) MSO-A and T3; 3) MSO-S and T3; 4) MSO-A and T7; 5) MSO-S and T7. K) 1 kb ladder.

C. elegans genomic blots.

C. elegans genomic DNA was separately digested with *Eco*RI and *Bam*HI, electrophoresed in a agarose gel and blotted to a nylon membrane. Human, sea urchin and *E. coli* (OP50) DNA were also used as controls. Two identical blots were prepared and individually hybridized to the end-labelled probes. Conditions for hybridization were as described in Materials and Methods except that the final washing was done at 42°C for 2 minutes and blots exposed for 3 days (Figure 7). Under these conditions no specific hybridization to *C. elegans* DNA could be detected. Moreover, no clear individual bands could be identified in the human and sea urchin DNAs (positive controls), indicating that the oligonucleotides were not specific enough to detect the homologous sequences in the DNA from these organisms. Also, hybridization bands could be detected in the *E. coli* DNA, which should not contain any homologous sequence.

3.2. Molecular characterization of the *dpy*-6 gene.

3.2.1. Tc1 transposon tagging the *dpy*-6 gene.

The *C. elegans* DR1013 is a mutator strain containing a copy of the Tc1 transposon disrupting the *dpy*-6 gene activity. This has been determined by complementation analysis, where the progeny arising from crossings between DR1013 males and dumpy hermaphrodites[6] exhibit a dumpy (non-unc) phenotype

---

[6] double mutants homozygous for the *dpy*-6(e14) reference allele and *unc*-20(e112).

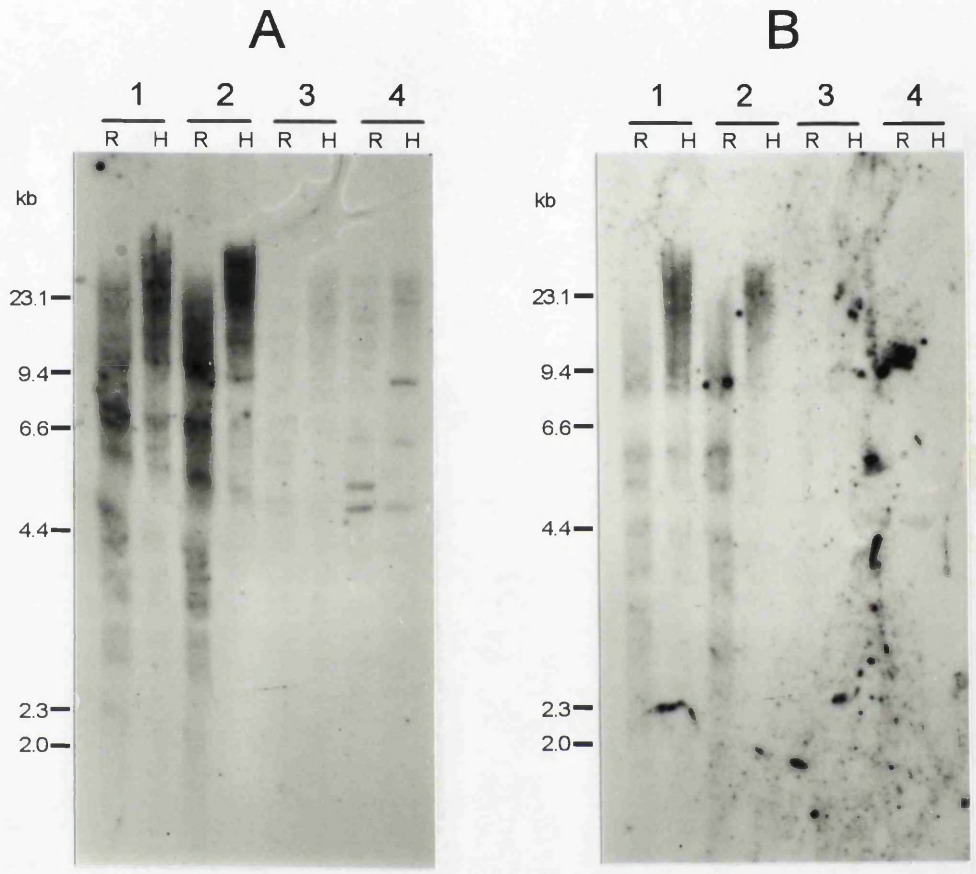


Figure 7. Southern blot analysis of MSO-S and MSO-A oligonucleotides hybridized to *C. elegans* DNA.

DNA was digested separately with *Eco*RI (R) or *Bam*HI (H), electrophoresed in a 0.7% agarose gel (1.5 V/cm for 20 h), blotted to a nylon membrane and hybridized to end-labelled MSO-S (panel A) or MSO-A (panel B) oligonucleotides.
1) human DNA (10 μg); 2) sea urchin DNA (5.4 μg); 3) *C. elegans* DNA (2.8 μg); 4) *E. coli* DNA (0.65 μg).

(P. Albert, personal communication). However, due to the presence of the *mut*-6 mutation this strain also has a much higher copy number of Tc1 transposable elements dispersed in its genome than the N2 strain. Tc1 transposition is derepressed in mutator strains, and Mori *et al.* (1988) suggested that the mutator locus itself could be a transposon. This is in agreement with current models for the regulation of transposition, where it is assumed that if a Tc element is inserted in the vicinity of an enhancer or promoter, the expression of the transposase could lead to a general derepression of transposition (van Luenen *et al.*, 1993). The high copy number of Tc1 transposons would make it difficult to identify the specific Tc1 insertion associated with the dumpy phenotype.

In order to reduce and stabilize the Tc1 copy number in the genome, a series of backcrosses to the wild-type N2 strain were performed, in an experimental approach identical to that previously described by Moerman and Waterston (1984) and Moerman *et al.* (1986). In this approach the Tc1-tagged allele is continually selected as the N2 chromosomal background becomes predominant. Considering that there is no positive selection for the *mut*-6 locus, when this allele is lost the transposition activity is reduced to the normal levels present in N2. Basically, two DR1013 hermaphrodites were independently crossed to N2 males and a single wild-type hermaphrodite from each progeny (F1) was selected. This hermaphrodite was then allowed to self-fertilize and a dumpy hermaphrodite from its progeny (F2) was selected and crossed to N2 males. This procedure was repeated six more times, then wild-type hermaphrodites (F2) were individually transferred to seeded

NGM plates. Two heterozygote hermaphrodites (one for each series of backcrosses) were selected and by self-fertilization used to originate a dumpy and a wild-type homozygous lineage. Figure 8 shows a schematic overview of the crosses involved in transposon tagging the *dpy*-6 gene.

The overall effect of the consecutive backcrosses was to remove the DR1013 chromosomal background, replacing it with N2 chromosomes while maintaining the *dpy*-6::Tc1 allele. After each round of backcross the Tc1 copy number should be reduced until reaching the number present in the N2 strain (approximately 30 copies per genome).

Four lineages were obtained using this procedure, resulting from the two independent series of backcrosses. The lineages were designated WT-A and DPY-A, arising from one series of backcrosses; and WT-C and DPY-C, arising from the other. Ideally, the dumpy lineage and its wild-type counterpart should be chromosomically identical, differing only by the extra Tc1 copy present in the dumpy lineage.

3.2.2. Identification of the Tc1 insertion associated with the dumpy phenotype.

The Tc1 insertion associated with the dumpy phenotype was identified by hybridization of a labelled Tc1 probe to a Southern blot containing *Eco*RI digested DNA of wild-type and dumpy lineages.

The probe was prepared by digesting pCE1003 (plasmid containing a cloned Tc1) with *Eco*RV and gel-purifying

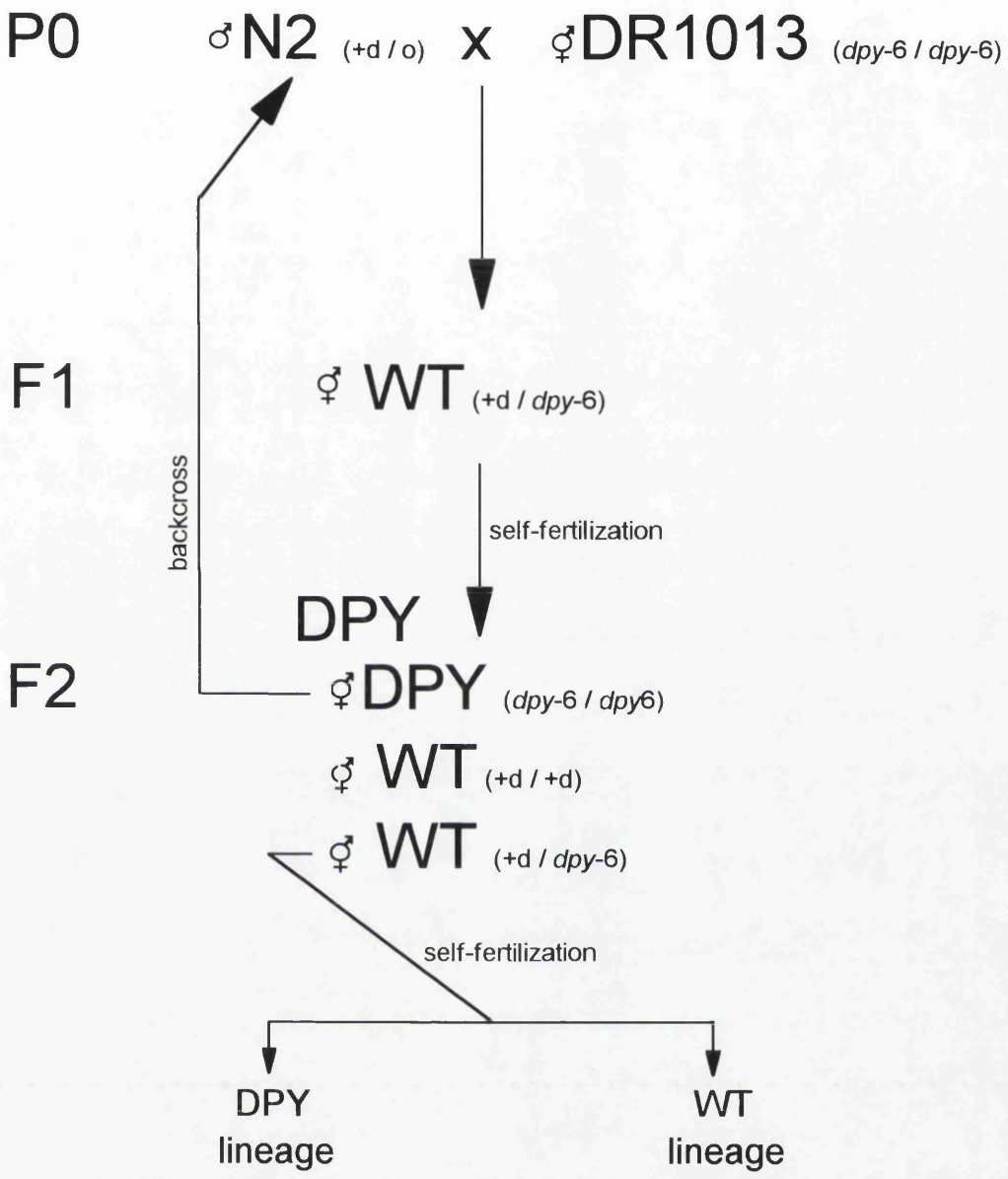


Figure 8. Schematic representation of the crossing/selfing procedures used for transposon tagging the *dpy*-6 gene.

The relevant alleles are:

+d - the wild-type *dpy*-6 allele;

dpy-6 - the mutant *dpy*-6::Tc1 allele.

Since the *dpy*-6 gene is located in the X chromosome and *C. elegans* males are XO, the wild-type male genotypes are represented by +d/O.

the 1.57 kb fragment containing Tc1. Tc1 has only two *Eco*RV sites located at 22 bp from the inverted repeat ends (Rosenzweig *et al.*, 1983), readily allowing the isolation of fragments containing almost all of Tc1, suitable for preparation as specific probes.

DNA from the N2, WT-A, WT-C, DPY-A and DPY-C strains was prepared, digested with *Eco*RI and electrophoresed in a agarose gel. The gel was subsequently blotted to a nylon membrane and hybridized to the ^{32}P-labelled Tc1 probe (Figure 9).

The hybridization pattern of the wild-type and dumpy congenic strains, as expected, is very similar to the N2 strain, indicating that the backcrosses were effective in removing almost all the extra copies of Tc1 present in the DR1013 strain. However, a few copies can be identified in the congenic strains that are not present in N2. Most of these extra copies of Tc1 are present in both wild-type and dumpy strains, indicating they are not associated with the dumpy phenotype. After the sixth round of backcrossing a fragment, approximately 9.5 kb in length, could be detected in the dumpy strains which was not present in the congenic wild-type animals. This fragment should contain the Tc1 insertion associated with the dumpy phenotype.

A second fragment, approximately 7.2 kb in length, was also detected exclusively in the dumpy strains (Figure 9B). However, when the hybridization pattern of DNA prepared from animals after the fourth round of backcross was analysed (Figure 9A), this extra fragment could also be detected in the WT-A strain, indicating that this Tc1 insertion was not linked with the dumpy phenotype. After two further backcrosses the

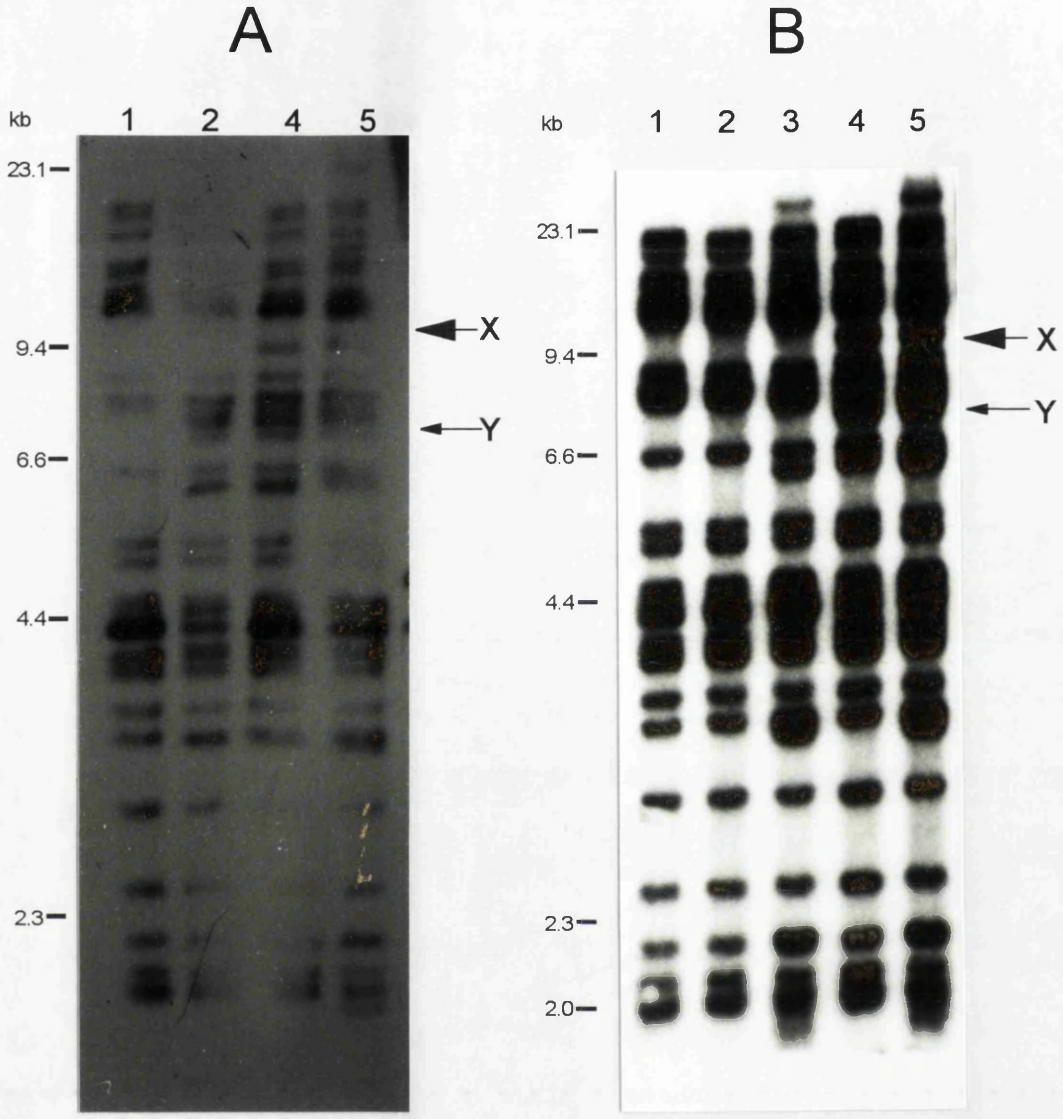


Figure 9. Tc1 hybridization to *Eco*RI digested DNA from wild-type and dumpy *C. elegans* strains.

2 μg of DNA were digested to completion with *Eco*RI and electrophoresed in a 0.6% agarose gel (2 V/cm, 22 h), then the gel was blotted to a nylon membrane and the filters hybridized to the ^{32}P-labelled Tc1 probe. Panel A corresponds to DNA extracted from worms after the fourth successive backcross and panel B corresponds to DNA extracted after the sixth series of backcross. The arrow (X) indicates the 9.5 kb fragment containing the Tc1 copy associated with the dumpy phenotype. The arrow (Y) indicates the 7.2 kb fragment containing another Tc1 copy.

1) N2; 2) WT-A; 3) WT-C; 4) DPY-A; 5) DPY-C.

Tc1 insertion was eliminated, entirely by chance, from the wild-type strains while being kept in the corresponding dumpy strains.

3.2.3. Molecular cloning of the Tc1-containing 9.5 kb fragment associated with the dumpy phenotype.

Since a single Tc1 insertion was found to be linked with the dumpy phenotype in the strains derived from DR1013, the first step towards characterizing the disrupted putative gene was cloning the corresponding DNA fragment carrying the Tc1 insertion. This was accomplished by cloning size selected DNA into the appropriate lambda vector and screening for the presence of Tc1.

Initial attempts were made using lambda vectors commonly used for genomic cloning, like λEMBL4 (Frischauf *et al.*, 1983) and λDASH II (Sorge, 1988). However, it was not possible to isolate any recombinants carrying the required fragment when both vectors were used, possibly due to the fragment size being very close to the minimum necessary for efficient packaging. Finally, λZAP II was used as the cloning vector, even though the fragment size was close to its upper limit.

3.2.3.1. Size selecting of DNA.

30 μg of DNA from strain DPY-C was digested to

completion with *Eco*RI and electrophoresed in a 0.8% LMP-agarose gel (2 V/cm, overnight) using a comb with wide teeth (4.5 cm) to form the wells. Using a scalpel blade, slices of 2-4 mm (approximately 1 ml volume) were excised from this gel and individually transferred to Eppendorf tubes. Ten slices were cut from parts of the gel containing DNA fragments ranging in size from 23 kb to 6 kb, melted at 65°C for 5 minutes and 30 µl samples run in an agarose gel together with a lane containing *Eco*RI digested DPY-C DNA (Figure 10A). In order to determine precisely the fractions containing the correct fragment, DNA from this gel was then transferred to a nylon membrane and hybridized to the ^{32}P-labelled Tc1 probe as before (Figure 10B).

Fractions four and five contained most of the 9.5 kb fragment. DNA from the remaining LMP-agarose corresponding to these fractions was then extracted and purified as described in Materials and Methods.

3.2.3.2. Cloning into λZAP II.

Approximately 250 ng of DNA from fractions four and five were separately ligated to 1 µg *Eco*RI digested CIP treated λZAP II DNA, in a 5 µl reaction. Half the ligation volume was then used for *in vitro* packaging. The library prepared using fraction four (λE4 library) contained $7.5x10^4$ pfu of phage and the library prepared using fraction five (λE5 library) contained $2x10^4$ pfu of phage. Applying the mathematical formula developed by Clarke and Carbon (1976), a *C. elegans* genomic library prepared by cloning 10 kb fragments and representative

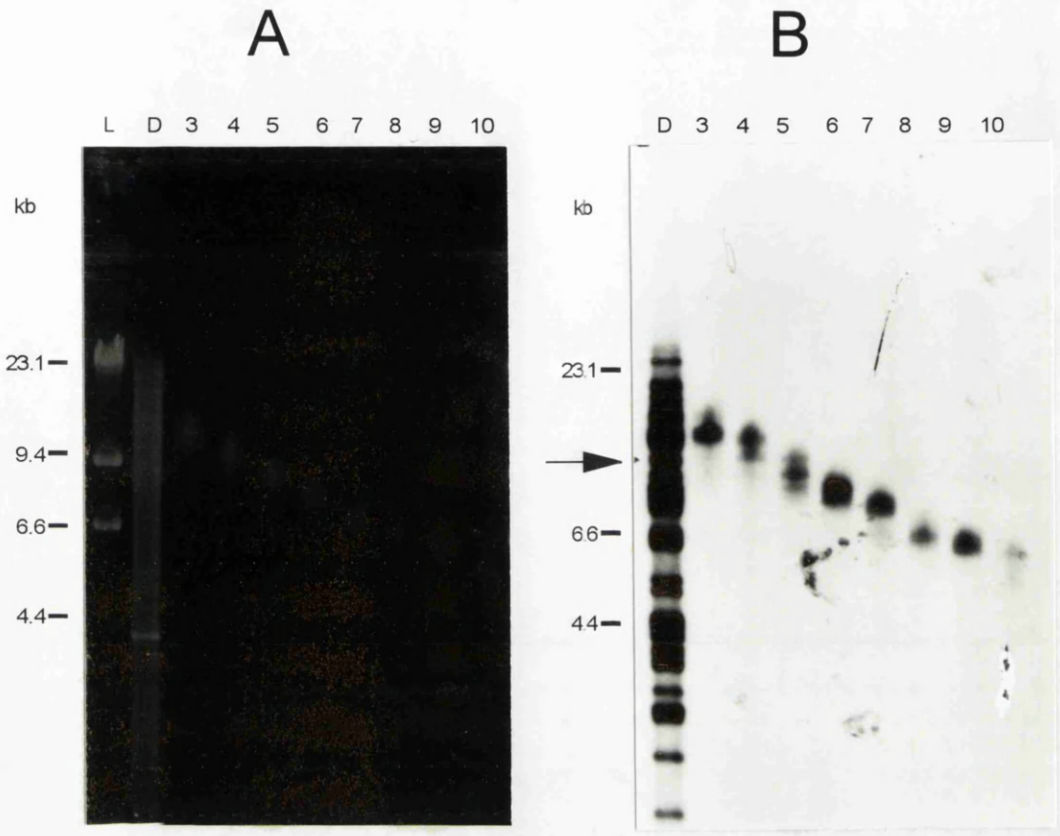


Figure 10. Agarose gel electrophoresis of *Eco*RI digested size fractionated DPY-C DNA and corresponding Tc1 hybridized Southern blot.

30 µl samples from melted LMP-agarose gel were run in a 0.6% agarose gel (panel A). The DNA was then blotted and hybridized to the ^{32}P-labelled Tc1 probe (panel B). The arrow indicates the location of the 9.5 kb fragment containing the Tc1 insertion associated with the dumpy phenotype.
L) λ/*Hind*III; D) DPY-C DNA *Eco*RI digested (800 ng); 3 to 10) different fractions from LMP-agarose.

of the entire genome (level of probability equal to 0.99) should contain at least $1x10^5$ recombinants. Considering that the cloned size fractionated DNA was probably less than 10% of the whole DNA content, both λE4 and λE5 libraries should contain at least one copy of the required DNA segment.

Fractions containing $1.5x10^4$ pfu of phage were plated from both libraries in 90 mm L-agar plates, the DNA was transferred to a nylon membrane and hybridized to the ^{32}P-labelled Tc1 probe. A single positive signal was detected in each screening. The plaques corresponding to these signals were collected, eluted in SM and rescreened at a lower phage density. Well isolated plaques from these secondary screenings were picked and used to prepare phage stocks.

DNA was prepared from both clones (designated λE4.2 and λE5.0, respectively), digested with *Eco*RI and electrophoresed in an agarose gel together with DPY-A and DPY-C DNA (Figure 11A). The insert contained in λE4.2 corresponded to the expected fragment size while λE5.0 contained a much smaller insert. In order to further verify that the insert in λE4.2 corresponded to the Tc1 tagged fragment associated with the dumpy phenotype, DNA from this gel was transferred to a nylon membrane and hybridized to the ^{32}P-labelled Tc1 probe (Figure 11B). The result from this Southern blot confirmed that λE4.2 carried a Tc1-containing insert that was also the same size as the Tc1 tagged fragment.

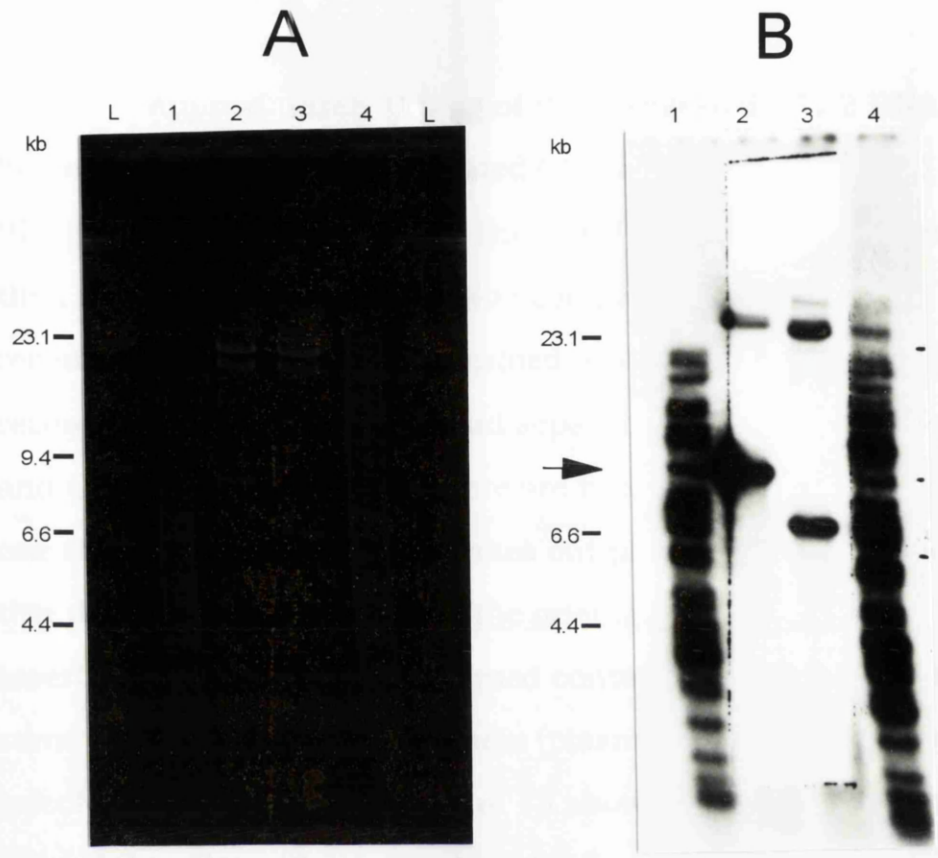


Figure 11. Agarose gel electrophoresis of *Eco*RI digested λE4.2 and λE5.0 DNA and corresponding Southern blot hybridized to Tc1 probe.

Approximately 500 ng of DNA from λE4.2 and λE5.0 was digested with *Eco*RI and electrophoresed in a 0.5% agarose gel, together with 2 μg of *Eco*RI digested DPY-A and DPY-C DNA (panel A). DNA was then transferred to a nylon membrane and hybridized to the ^{32}P-labelled Tc1 probe (panel B). The arrow indicates the location of the 9.5 kb fragment containing the Tc1 insertion associated with the dumpy phenotype. The hybridization bands larger than 20 kb correspond to partial digestions of phage DNA.

L) λ/*Hind*III; 1) DPY-A; 2) λE4.2; 3) λE5.0; 4) DPY-C.

3.2.3.3. Subcloning into pBluescript II.

Approximately 0.6 μg of *Eco*RI digested λE4.2 DNA was ligated to 0.3 μg of *Eco*RI digested CIP-treated pBluescript II (KS-) DNA in a 10 μl reaction. One third from this reaction was then used to transform XL1-Blue competent cells. Sixteen recombinant colonies were obtained. DNA from some of these recombinants was prepared and separately digested with *Eco*RV and *Cla*I (Figure 12). Since there are two *Eco*RV sites and only one *Cla*I site in Tc1 (both enzymes cut pBluescript only once), this digestions would confirm the orientation of the cloned inserts. All recombinants analysed contained the insert in the same orientation and one of them (plasmid preparation 2) was selected for further work. Figure 13 shows a partial restriction map of this plasmid, designated pB9.5.

Two other plasmids derived from pB9.5 were constructed in order to facilitate sequencing the DNA flanking both sides of Tc1 in the cloned fragment. DNA from pB9.5 was digested with *Eco*RV and the fragments separated by electrophoresis in LMP-agarose. The bands containing the 7 kb (pBluescript plus the right insert) and 3.6 kb (left insert from pB9.5) fragments (see Figures 12A and 13), were then excised and the DNA was purified. DNA from the 7 kb fragment was religated and used to transform XL1-Blue competent cells. Recombinant (white) colonies were selected, the DNA was prepared, digested with *Eco*RV and analysed by agarose gel electrophoresis. One plasmid was selected for further work (pBlueR, Figure 14).

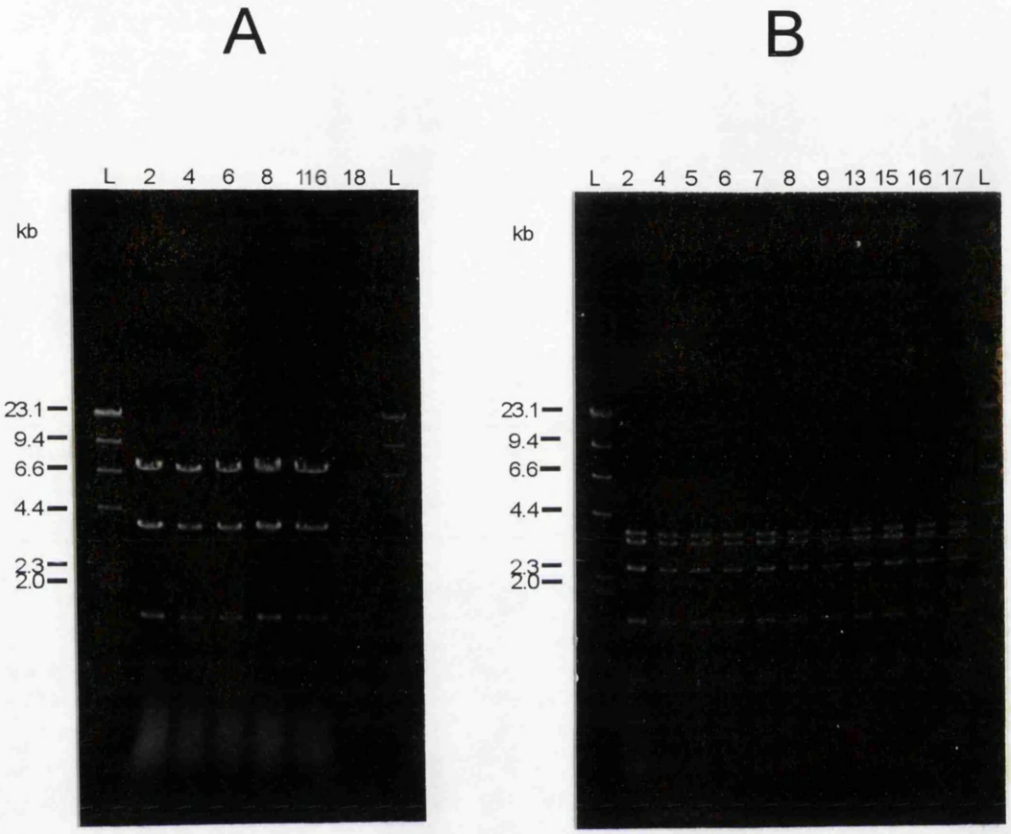


Figure 12. Agarose gel electrophoresis of DNA from pBluescript recombinants digested with *Eco*RV and *Cla*I.

300 ng of plasmid DNA was digested with *Eco*RV (panel A) or *Cla*I (panel B) and electrophoresed in a 0.7% agarose gel. The bands in panel A are: 7 kb, pBluescript plus the right insert from pB9.5; 3.6 kb, left insert from pB9.5; 1.57 kb, Tc1 transposon (excluding 22 bp from each inverted repeat).
L) λ/*Hind*III; 2 to 18) DNA from different recombinant plasmids.

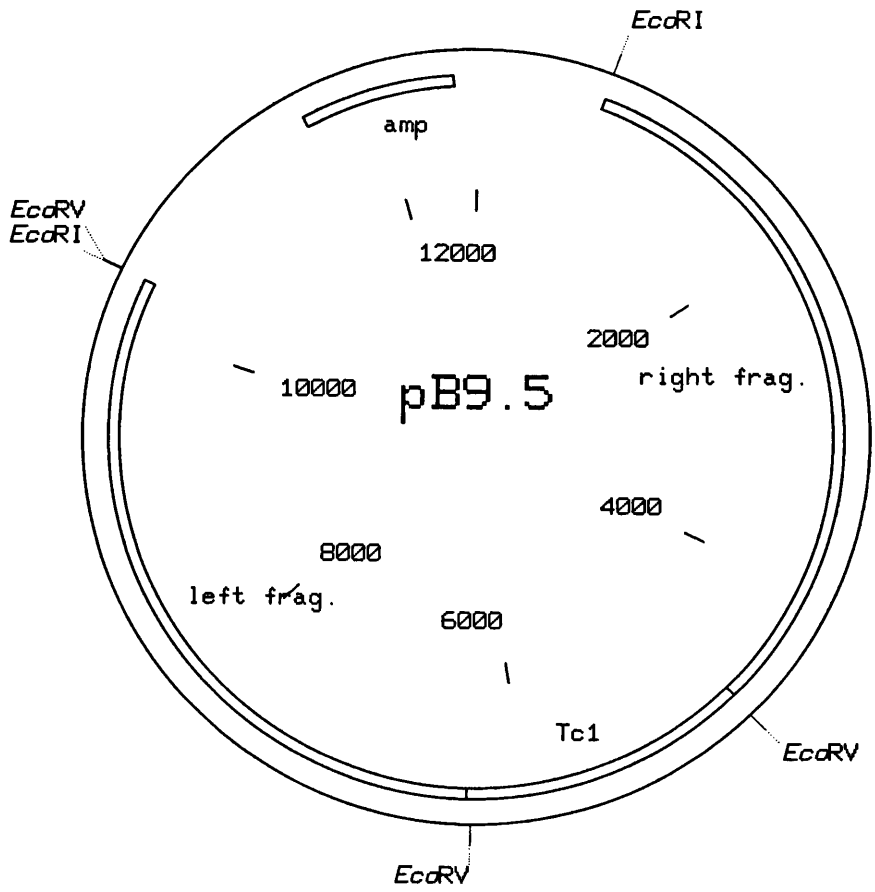


Figure 13. Partial restriction map of plasmid pB9.5.

The position of the 9.5 kb fragment is indicated together with the Tc1 insertion. The Tc1 sequence divides the insert into two distinct sequences, marked **right frag** and **left frag**. Also indicated is the ampicillin resistance gene (amp) from pBluescript. Only the relevant restriction enzyme sites are shown.

DNA from the 3.6 kb fragment was ligated to *Eco*RV digested and dephosphorylated pBluescript SK+ and used to transform XL1-Blue competent cells. The transformants included white, blue and pale blue colonies. DNA was prepared from some of the white and pale blue colonies and the size and orientation of the insert was verified by digestion with *Eco*RI followed by agarose gel electrophoresis. All white colonies contained different plasmids of a smaller size than expected for the correct construct. Two of the pale blue colonies contained the plasmid carrying the 3.6 kb fragment in the desired orientation (with the 22 bp from the Tc1 sequence closer to the SK primer site)(pBlueL, Figure 14).

The plasmids pBlueR and pBlueL would allow sequencing from the KS primer through the cloned inserts, including the 22 bp from Tc1 ends, thus enabling the determination of the Tc1 insertion site as well the flanking sequences immediately adjacent.

3.2.4. Location of the 9.5 kb Tc1 tagged insert in the physical map of the *C. elegans* genome.

The location of the corresponding wild-type allele of the tagged gene was determined by hybridizing a specific probe from the cloned 9.5 kb fragment to a collection of overlapping cosmid or YAC clones representing most of the *C. elegans* genome.

Approximately 90-95% of the *C. elegans* genome has been cloned into more than 17000 cosmids assembled into some 700 clusters (Coulson *et al.*, 1986). These clusters (contigs) have

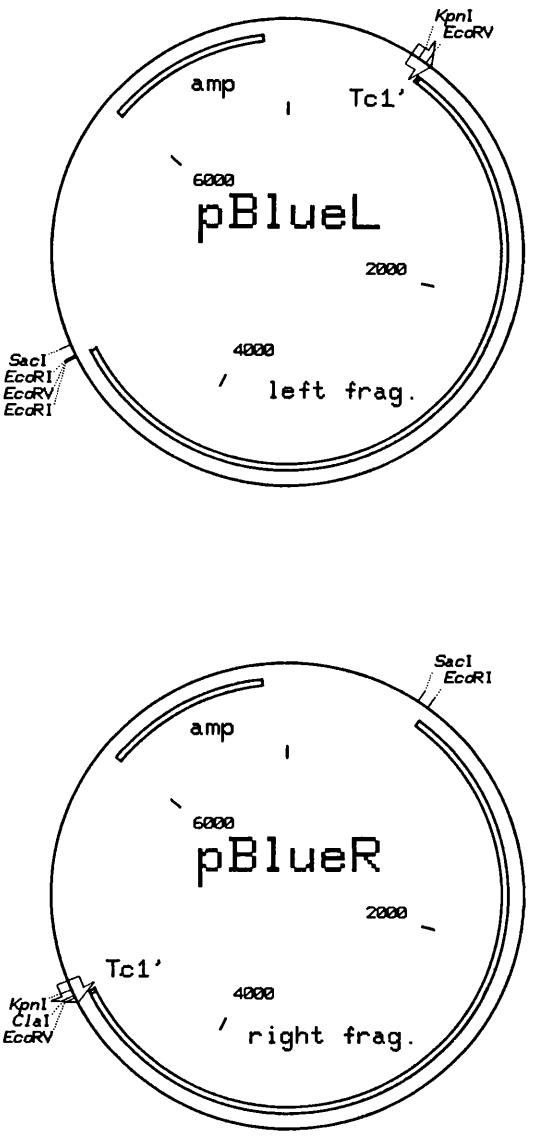


Figure 14. Partial restriction maps of plasmids pBlueR and pBlueL.

The arrow indicates the location of the KS sequencing primer. Tc1' corresponds to the 22 bp sequence from Tc1 ends. The relative sizes of both plasmids are shown proportional to pB9.5 (Figure 13).

been further linked by cloning large (50 to 1000 kb) genomic fragments into a YAC vector[7] (Coulson *et al.*, 1988). This collection of overlapping clones constitutes a physical map of the *C. elegans* genome and since many cloned genes had previously been assigned a position in the genetic map, the correlation between maps can be established.

DNA from pBlueL was digested with *Eco*RV, electrophoresed in a LMP-agarose gel and the band corresponding to the 3.6 kb fragment was excised and purified from the gel. DNA was then labelled and hybridized to a grid containing DNA from 958 different YACs[8].

Three positive hybridization signals were found, corresponding to YACs Y51E1, Y75D9 and Y58B1 (Figure 15). The cloned inserts from these three YACs overlap considerably and their location in the physical map (Figure 16) correspond to a region in the chromosome X of *C. elegans* where the *dpy*-6 gene had previously been mapped by three-factor crosses (Brenner, 1973) (Figure 17).

In order to identify the cosmids carrying the wild-type allele corresponding to the Tc1 tagged 9.5 kb fragment, nine cosmids covering the entire overlapping region present in the YACs were selected[9]. Bacterial colonies carrying the selected

---

[7] YAC vectors (Burke *et al.*, 1987) provide centromeric, telomeric and selective functions for the constructs, which are then introduced into *Saccharomyces cerevisiae* and replicate in the same manner as the host chromosomes.

[8] a gift from A. Coulson (MRC-Cambridge).

[9] the selected cosmids were: T13E6, K01F10, C07A6, K04A4, F37F1, B0420, F16F9, W04H3 and T01B10 (see Figure 16A for their relative locations).

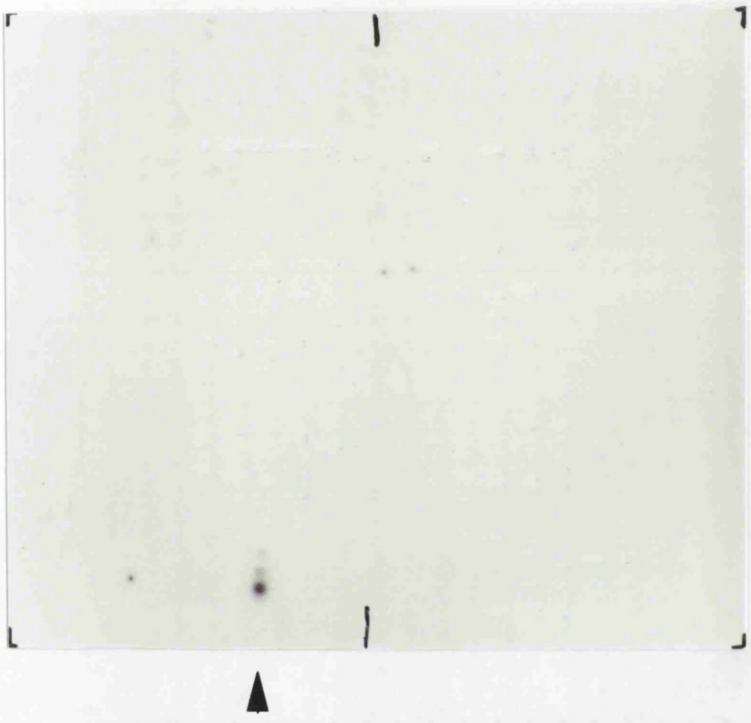


Figure 15. Hybridization of the insert from pBlueL to a grid of YACs containing cloned genomic *C. elegans* sequences.

DNA from pBlueL insert was purified, labelled by random priming and hybridized to a grid containing 24 rows by 40 columns of yeast colonies containing YACs. The arrow indicates the location of the three positive signals. Other spots of photographic deposit do not correspond in position to yeast colonies.

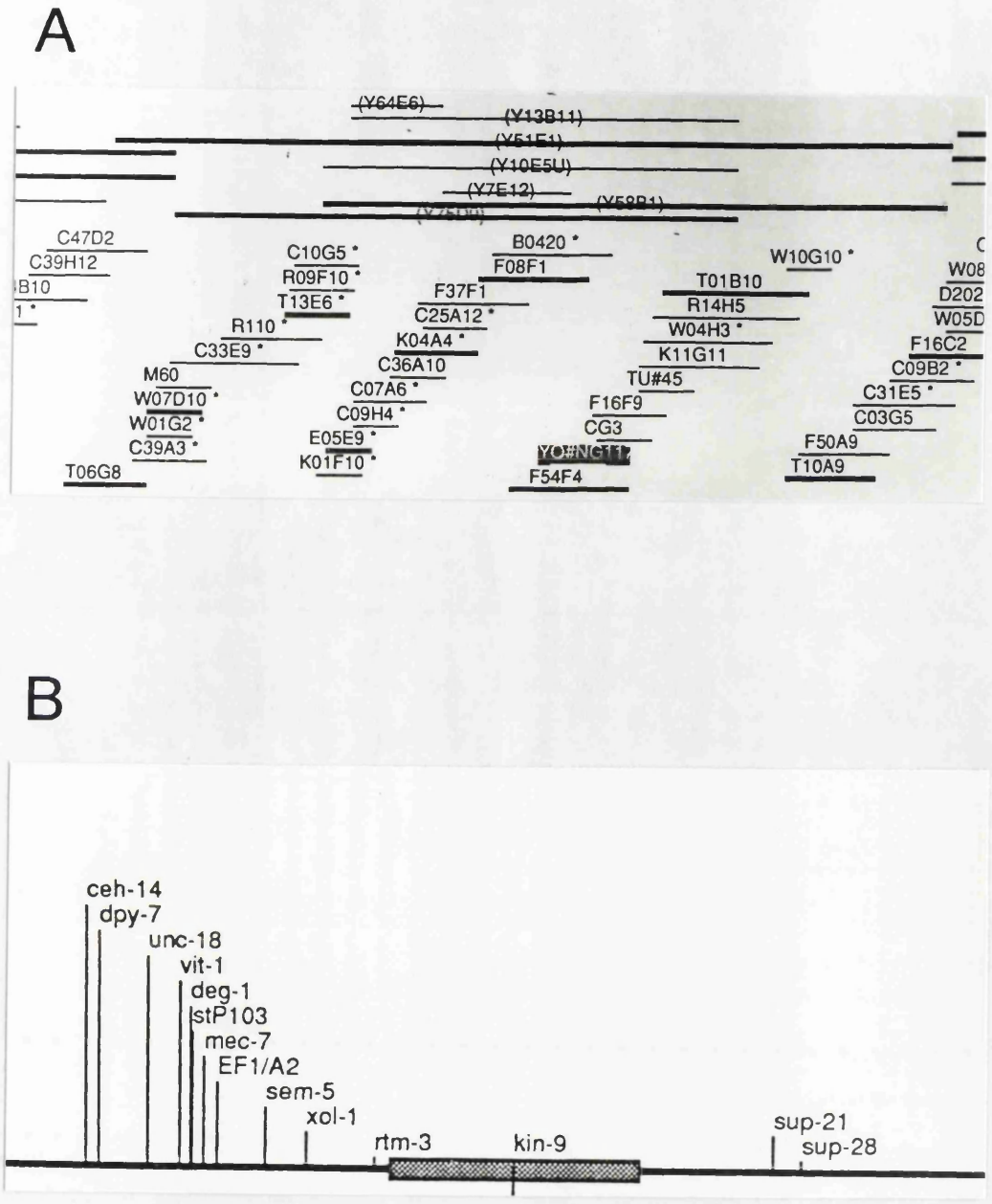


Figure 16. Physical map from part of *C. elegans* chromosome X.

The lines in panel A indicate the relative positions of YAC and cosmid clones (long and short lines in the upper and lower parts of panel A, respectively). Panel B presents a more general view of this region from chromosome X with the relative position of cloned genes indicated. The box indicates the region which is shown in panel A. Data made available by the Caenorhabditis Genetic Center.

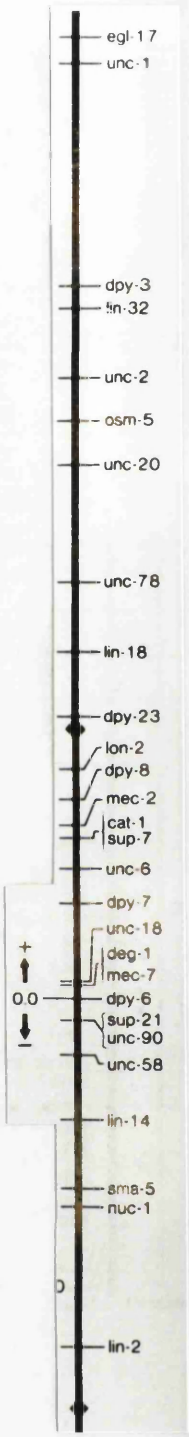


Figure 17. Genetic map from part of *C. elegans* chromosome X.

The map indicates the relative position of genes mapped by two-factor and three-factor crosses. The scale of the map is 1 map unit to 0.6 cm (1 map unit is equivalent to 1% recombination). *dpy*-6 is located at position 0.0.Data made available by the Caenorhabditis Genetic Center.

cosmids were plated in L-agar plates containing the appropriate antibiotic. Individual colonies, ten for each cosmid, were then screened by colony hybridization using the ^{32}P-labelled 3.6 kb insert from pBlueL as a probe (Figure 18). Hybridization was detected to colonies carrying the cosmids B0420, F16F9 and W04H3[10]. Some individual colonies did not hybridize to the probe, presumably due to cosmid deletion during colony growth.

DNA from these three cosmids was prepared, digested with *Eco*RI and analysed by agarose gel electrophoresis (together with *Eco*RI digested genomic DNA from N2, WT-A and WT-C strains) (Figure 19A). Table 1 shows the corresponding fragment sizes from the *Eco*RI digested cosmids. DNA from this gel was then transferred to a nylon membrane and probed with the ^{32}P-labelled insert from pBlueL (Figure 19B). Hybridization was detected to one fragment from each cosmid, with sizes corresponding to 7.36 kb (B0420), 7.88 kb (F16F9) and 3.79 kb (W04H3). The hybridizing fragment from cosmid F16F9 is the same size as those detected in the genomic DNA of wild-type strains (Figure 19B, lanes 7-9), indicating that this cosmid contained the wild-type counterpart of the cloned 9.5 kb fragment tagged with Tc1. Based on the *Eco*RI fragment sizes and the hybridization pattern detected (Figure 19), and taking into account the mapping data available from the physical map (Figure 16) a partial restriction map including the relevant DNA fragments was established (Figure 20).

---

[10] The cosmid B0420 has its insert cloned into pJB8 (Ish-Horowicz and Burke, 1981) whilst cosmids F16F9 and W04H3 have theirs cloned into LoristB (Cross and Little, 1986).

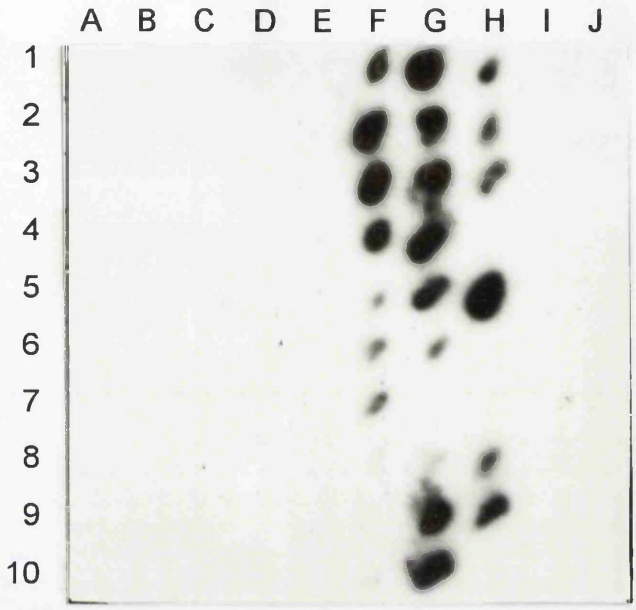


Figure 18. Hybridization of the insert from pBlueL to DNA from colonies carrying cosmids covering the overlapping region of YACs Y51E1, Y75D9 and Y58B1.

Colonies were transferred to a Hybond-N membrane placed on the surface of an L-agar plate and allowed to grow at 37°C for 4 h. The membrane was removed, the colonies lysed and the DNA denatured and subsequently neutralized. Following this, the DNA on the membrane was hybridized to the ^{32}P-labelled insert from pBlueL.

A) T13E6, B) K01F10, C) C07A6, D) K04A4, E) F37F1, F) B0420, G) F16F9, H) W04H3, I) T01B10.

1 to 10) different colonies containing copies of the same cosmid.

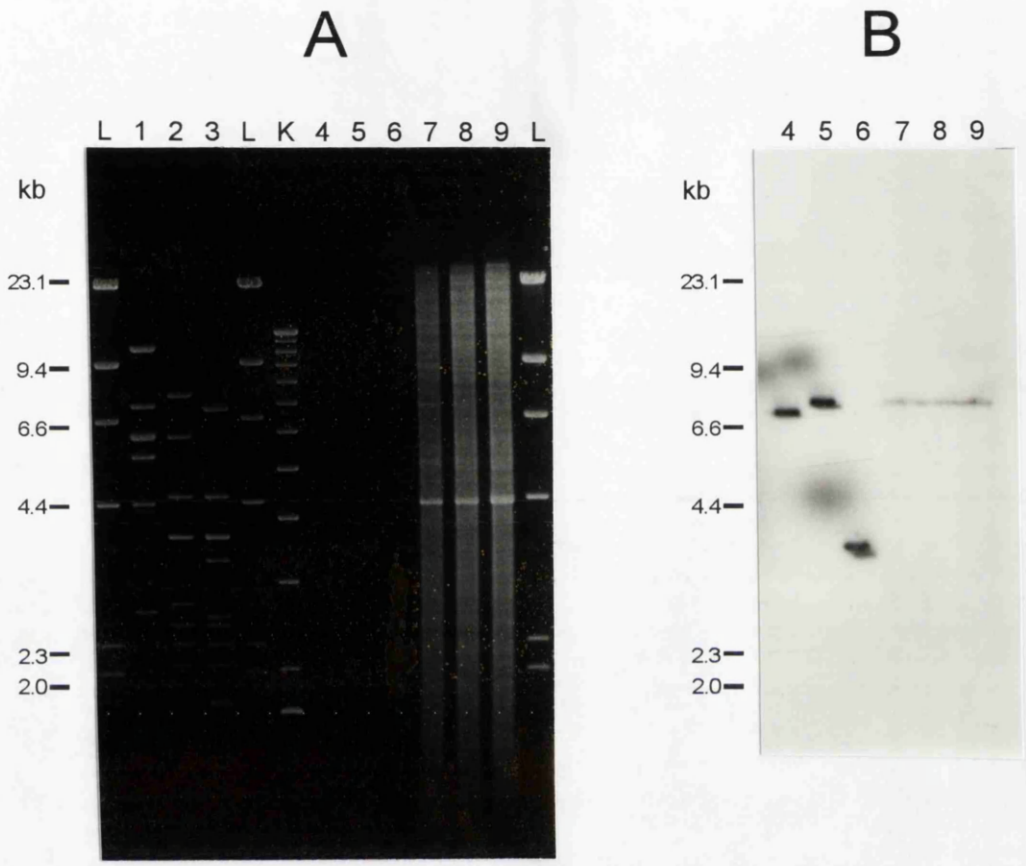


Figure 19. Agarose gel electrophoresis of DNA from cosmids B0420, F16F9 and W04H3 digested with *Eco*RI and the corresponding Southern blot hybridized to ^{32}P-labelled insert from pBlueL.

DNA was digested with *Eco*RI and electrophoresed in a 0.7% agarose gel (2 V/cm, overnight)(panel A). DNA was then transferred to a nylon membrane and hybridized to the insert from pBlueL (panel B).
K) 1 kb ladder; L) λ/*Hind*III; 1) B0420 (≅0.7 μg); 2) F16F9 (≅0.7 μg); 3) W04H3 (≅0.7 μg); 4) B0420 (≅1.7 ng); 5) F16F9 (≅1.7 ng); 6) W04H3 (≅1.7 ng); 7) N2 (2 μg); 8) WT-A (2 μg); 9) WT-C (2 μg).

1	2	3
10.74		
	7.88*	
7.36*		
		7.20
6.21	6.21	
5.61		
	4.63	4.63
4.42		
	3.79	**3.79***
		3.38
	2.76	
2.67		
		2.62
	2.51	2.51
	2.30	2.30
	2.09	2.09
		1.71
	1.50	1.50
1.48		
	1.38	
	1.26	1.26
		1.19
		1.09
0.94	0.94	
0.83	0.83	
46.48	41.88	40.79

Table 1. Size of the *Eco*RI fragments from cosmids B0420, F16F9 and W04H3.

The sizes (in kb) were calculated based on the gel from Figure 19A and are shown following the layout of that gel. Numbers in bold correspond to bands containing two fragments of the same size (as estimated from band intensities). Numbers marked with (*) correspond to fragments hybridizing to the insert from pBlueL. 1) B0420; 2) F16F9; 3) W04H3.

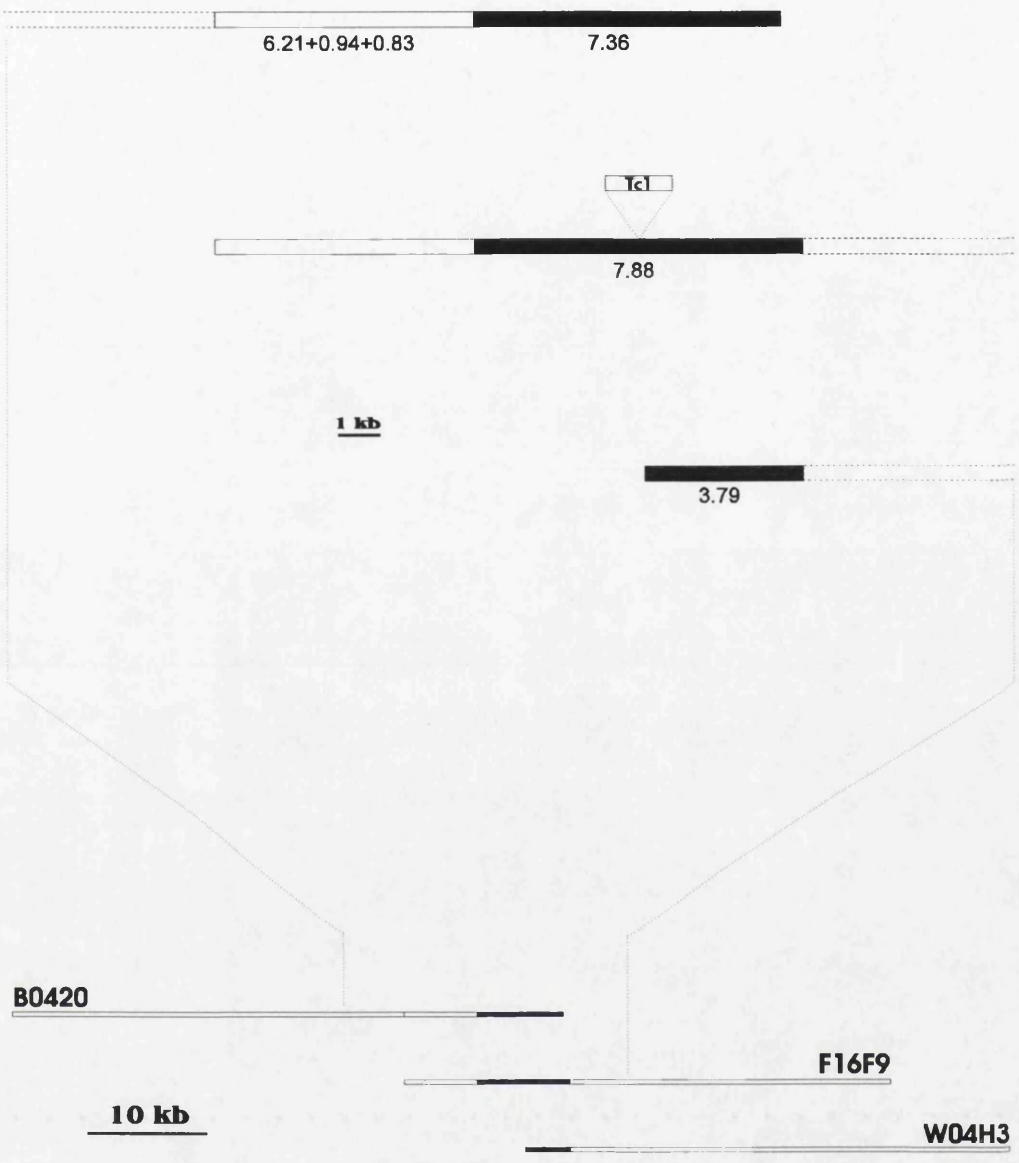


Figure 20. Partial restriction maps of cosmids B0420, F16F9 and W04H3.

The *Eco*RI fragments hybridizing to the insert from pBlueL are marked in black. An enlarged view of the overlapping region from the three cosmids is shown at the top of the figure. Sizes (in kb) are indicated below the fragments. The approximate location of the Tc1 insertion present in the dumpy strains derived from DR1013 is also indicated.

Considering that Tc1 is 1.61 kb in length, its insertion in a 7.9 kb *Eco*RI genomic fragment would increase the size of this fragment to 9.5 kb, which is the size detected in the dumpy strains. This is shown in Figure 21, where genomic DNA from N2, WT-A, WT-C, DPY-A and DPY-C strains, together with λE4.2 DNA, was digested with *Eco*RI, electrophoresed in an agarose gel and the DNA blotted to a nylon membrane. DNA on this membrane was then hybridized to the ^{32}P-labelled insert from pBlueL.

3.2.5. Attempts to rescue the dumpy phenotype.

Since the cosmid F16F9 contained the 7.9 kb *Eco*RI fragment equivalent to the 9.5 kb fragment tagged in the dumpy strains derived from DR1013, attempts were made to rescue the dumpy phenotype using this and the other two contiguous cosmids (B0420 and W04H3). This would provide confirmation that the wild-type sequence was the *dpy*-6 gene.

Two sets of microinjections were performed. In the first set, DNA from each cosmid was separately injected into CB14 hermaphrodites. Five animals were injected with each cosmid. In the second set, a mixture of all three cosmids was used for the microinjections (ten animals were injected). These same two sets of microinjections were repeated in a control experiment using DPY-C hermaphrodites. No wild-type transgenic nematodes were found amongst the F1 progeny arising from these animals (more than 100 animals were analysed for each cosmid injected).

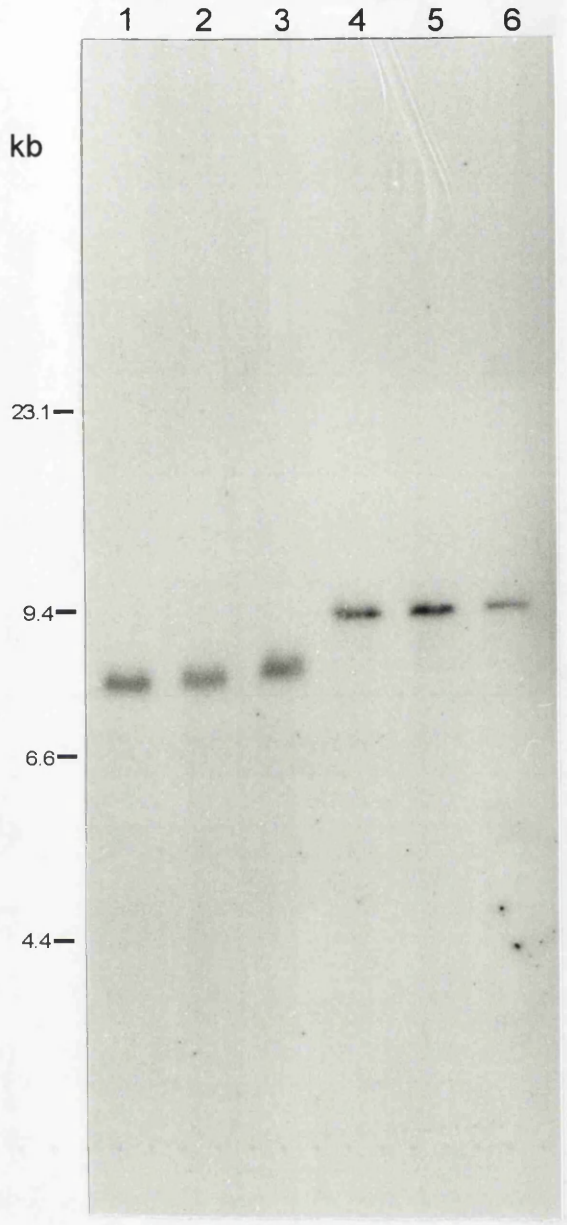


Figure 21. Southern blot analysis of *Eco*RI digested DNA from wild-type and dumpy strains and λE4.2 hybridized to the ^{32}P-labelled insert from pBlueL.

DNA was digested with *Eco*RI and electrophoresed in a 0.5% agarose gel (2V/cm, overnight). DNA was then blotted to a nylon membrane and hybridized to the ^{32}P-labelled insert from pBlueL.

1) N2 DNA (2 μg); 2) WT-A DNA (2 μg); 3) WT-C DNA (2 μg); 4) DPY-A DNA (2 μg); 5) DPY-C DNA (2 μg); 6) λE4.2 (180 pg).

3.2.6. Sequencing the cloned inserts from pBlueR and pBlueL.

ssDNA was prepared from these plasmids and sequencing was initially performed using the KS primer. As shown in Figure 22, the exact Tc1 insertion point could then be determined. Tc1 is invariably inserted adjacent to a TA dinucleotide, which is duplicated in consequence of the insertion (Emmons, 1988; H. van Luenen, C. Vos and R. Plasterk, personal communication).

Sequencing then progressed by designing oligonucleotide primers complementary to the ends of the previously sequenced stretch of DNA. Using this sequencing strategy, a total of 2880 bp were determined (1427 bp from pBlueL and 1453 bp from pBlueR). In order to assemble a continuous sequence excluding the Tc1 insertion, the sequence data from pBlueL were reversed and complemented, the duplicated TA dinucleotide was removed and joined to the sequence determined from pBlueR.

3.2.7. Sequence analysis.

The sequence contains an ORF with 1857 bp (extending from position 193 to 2049), capable of coding for a protein containing 619 amino acids. When this sequence was analysed with programmes that predict potential coding regions (Staden, 1984bc; Fickett, 1982), the ORF appeared likely to be a coding sequence (Figure 23).

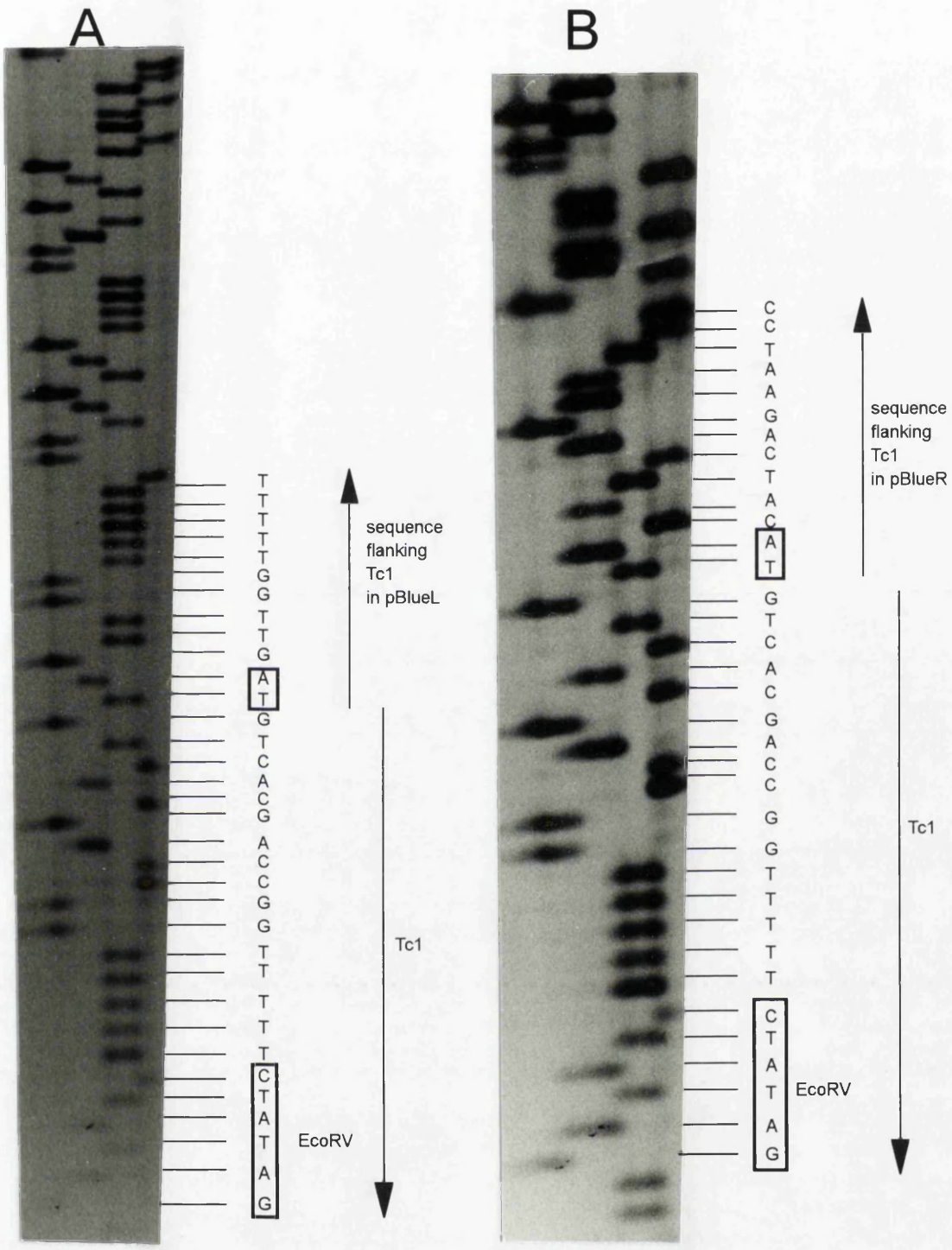


Figure 22. Sequence of pBlueL and pBlueR near the Tc1 insertion point.

ssDNA was prepared from both plasmids and sequenced using the KS primer. The sequence corresponding to Tc1 is indicated as well as the TA dinucleotide duplicated at the Tc1 insertion point.

Panel A: pBlueL; Panel B: pBlueR.

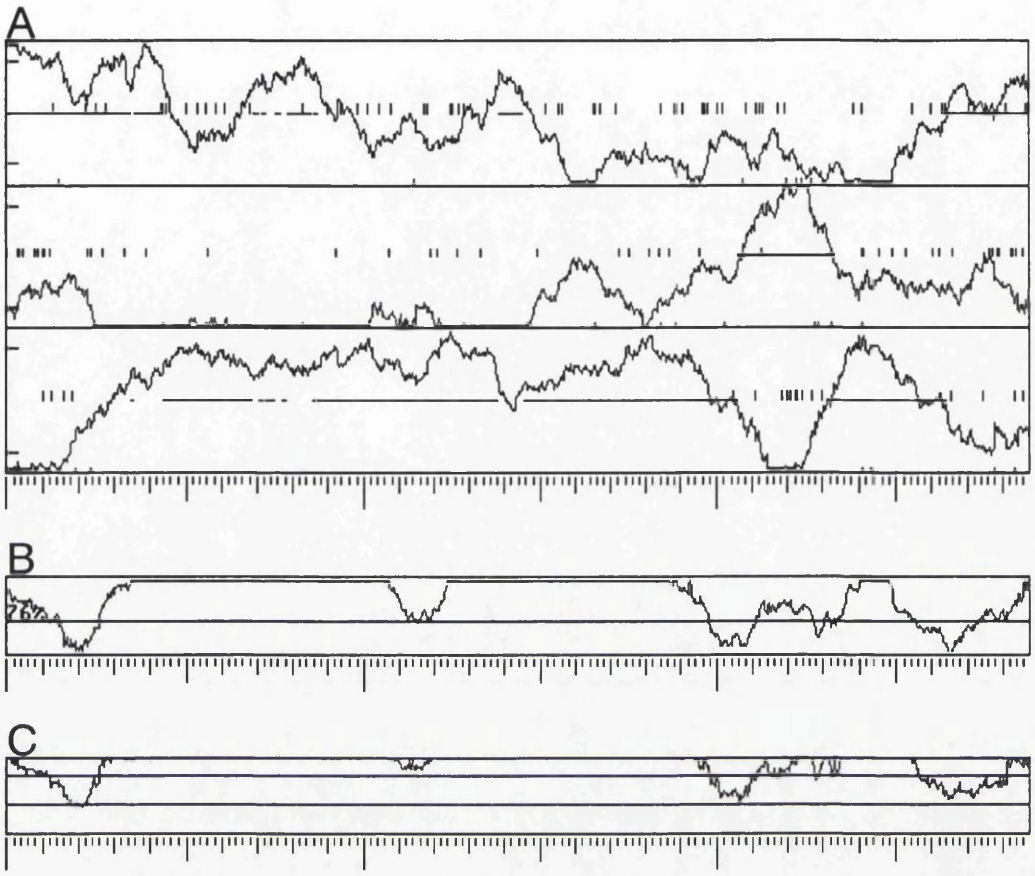


Figure 23. DNA sequence analysis.

Potential protein-coding sequences were determined by three different methods. The graph in panel A was plotted using the Positional Base Preference method (Staden, 1984b). The method takes advantage of the uneven use of amino acids by proteins and the structure of the genetic code table and assumes there is a typical amino acid composition and no codon preference. Each reading frame of the sequence is represented by a rectangle (frame I in the bottom). Peaks result when a particular frame gives consistently the highest scores. The vertical marks in the middle of each rectangle represent non-sense codons. The vertical marks at the bottom of each rectangle represent methionine codons. The graph in panel B is a Uneven Positional Base Frequencies plot (Staden, 1984b). The bar drawn across the plot corresponds to a level that is exceeded by 76% of known coding sequences but is reached by only 24% of known non-coding sequences. Panel C corresponds to a Fickett testcode graph (Fickett, 1982) and is plotted relative to three levels of decision: top division corresponds to "coding", middle division corresponds to "no opinion" and the bottom division corresponds to "non-coding". Each division in the scales beneath the graphs corresponds to 20 bp.

The complete nucleotide sequence and the corresponding amino acid sequence translated from the largest ORF is presented in Figure 24. This sequence has some features compatible with coding sequences. Just upstream of the start codon there are sequences similar to vitellogenin promoter elements described in *C. elegans* (Spieth *et al.*, 1985; MacMorris *et al.*, 1992). One of these elements, identical to VPE1 (*vit* promoter element 1)[11], has the consensus sequence TGTCAAT and is located 54 bp upstream of the putative starting codon. The second site, similar to the TATA box (Corden *et al.*, 1980), is located 41 bp from the starting codon. Two sequences very similar to the consensus for donor and acceptor splice sites (Emmons, 1988) are found starting at positions 1069 and 1252. The *C. elegans* donor consensus sequence ([A/G]AG'GTAAGTT) is nearly identical to the general eukaryote consensus (Mount, 1982) whilst the acceptor sequence (TTTCAG'[G/A]) differs more extensively[12]. These sites delimit a putative intron containing 186 bp. Significantly, the coding sequence prediction scores are lower in this region. Sequences homologous to the polyadenylation signal can be found downstream of the termination codon. One of these sequences, starting at position 2838, is identical to the polyadenylation signal functional in other eukaryotes (Proudfoot and Brownlee, 1976). Other putative polyadenylation sites that could be functional in *C. elegans* (T. Blumenthal, O. White and C. Fields; personal

---

[11] this element is similar to the mammalian CCAAT box.

[12] the eukaryote consensus sequences for donor and acceptor splice sites are (C/A)AG'GT(A/G)AGT and $(T/C)_{11}$N(C/T)AG'G, respectively.

```
CTGTTAAAGTTACACCAAAATTGGAGCTTTCATTTGATGAGCCAACTGAA      50

ATCACAAAAGCTCCACATCCAGGCAAGTTGCTTGAAAAGTTTAAAACCAA     100

TAATTTAATTAACTTTTTACATGAATTAATTTAAAAATTGTCAATGGACA     150

CTATATTTCTAACGGGAGAACGTGTCTCAAAGATAGAAAAATATGTATTC     200
                                          M  Y  S        3

ACATTTATTATCAATACTTATTAGTTTGAATCAATTGATGAGAAGGTCTT     250
 H  L  L  S  I  L  I  S  L  N  Q  L  M  R  R  S  F     20

TTGAAGGAGTCGAATTTATAGTAGATTTTAAAGAAAAACATATCATTTTG     300
  E  G  V  E  F  I  V  D  F  K  E  K  H  I  I  L       36

TTCTTTCGGACAATTTTGCACGTTACAGTGAAGCCAAGGAAAACGACGAC     350
F  F  R  T  I  L  H  V  T  V  K  P  R  K  T  T  T      53

TACAACCACCTCGACTACAACCACTACAGAGAAGCCAGTAATTACAGAGG     400
 T  T  T  S  T  T  T  T  T  E  K  P  V  I  T  E  G     70

GGTCGACCACAAACAAGCAGCCAACAACATCTGAACATGAATCTACCACT     450
  S  T  T  N  K  Q  P  T  T  S  E  H  E  S  T  T       86

GAGGTACCGACAACTACGGAAGAGCCAACCACCACGACCGGACAACACAA     500
E  V  P  T  T  T  E  E  P  T  T  T  T  G  Q  H  K     103

GACTGAGAAGGCAATTACTACAGAGGAACCAACTACTACTGAAGAATCAT     550
 T  E  K  A  I  T  T  E  E  P  T  T  T  E  E  S  S    120

CCACAACTGAAGAAGTAACTACCACAGAGGAGCCAGCTAACACTGGAAAT     600
  T  T  E  E  V  T  T  T  E  E  P  A  N  T  G  N      136

CCGCCCACAACTGAAAACCCAACAACCACAGAACAGCCAACCAGTACTGC     650
P  P  T  T  E  N  P  T  T  T  E  Q  P  T  S  T  A     153

CGAGTCAACTACCACGGCACTACCATTCACGACTGAGCAGACAGTTACCA     700
 E  S  T  T  T  A  L  P  F  T  T  E  Q  T  V  T  T    170

CAGAGGAACCAACAACTGCGGAAAAGTCAACCGCTACTCAGAAGCCAACA     750
  E  E  P  T  T  A  E  K  S  T  A  T  Q  K  P  T      186
```

```
ACCACACAGGAGTCAGTTTCCACGGAAAAGACATCCACTACCAAAAAGGC    800
T  T  Q  E  S  V  S  T  E  K  T  S  T  T  K  K  A     203

ATCCACCACAGAGGAGCCAACCACTACTGATGAGCCAACCACCACGACAG    850
 S  T  T  E  E  P  T  T  T  D  E  P  T  T  T  T  E    220

AGTCATCCACAACCGGCAAGGCAACAACCCCGGAGTTGTCAACTACTTCT    900
  S  S  T  T  G  K  A  T  T  P  E  L  S  T  T  S     236

GAAGAGACGACCACCACGGAATTGAAAATCACTACTGAAGGTTCAACGAC    950
E  E  T  T  T  T  E  L  K  I  T  T  E  G  S  T  T     253

TACAGAAGAGCCAACCACTACAGCTATTTTCGCTGAAGCTTCAACCGGCA   1000
 T  E  E  P  T  T  T  A  I  F  A  E  A  S  T  G  I    270

TTATTACAACTGACGAGGAAACAACATCAACCACCTCTACCACACCTGAG   1050
  I  T  T  D  E  E  T  T  S  T  T  S  T  T  P  E     286

ATCACGTCTACAAAAGAAATCgtaactgaatctgcaatcactcaaacatc   1100
I  T  S  T  K  E  I  v  t  e  s  a  i  t  q  t  s     303

cgtttctgtggtcgaaagttcaaccccacgccaactgccagaaagatgga   1150
 v  s  v  v  e  s  s  t  p  r  q  l  p  e  r  w  k    320

aggcgattgtcaacaagtttaaacacaatcttgaagtactaaaggaaaag   1200
  a  i  v  n  k  f  k  h  n  l  e  v  l  k  e  k     336

aaacgactgttgaaagaaaaagagagcacgtcgactactggatcagatag   1250
k  r  l  l  k  e  k  e  s  t  s  t  t  g  s  d  s     353

ttctgagACAACAACTGTAGTTGCTGAAAACATCGATGAGGTTACTACGA   1300
 s  e  T  T  T  V  V  A  E  N  I  D  E  V  T  T  T    370

CCGAGAAAGAGAAAGTTGTTCAGACAACACCAATAACAACAGAAAAATCG   1350
  E  K  E  K  V  V  Q  T  T  P  I  T  T  E  K  S     386

ACAACACAGGAAGAGACTACAACAACAACAACCACCACGGAGAAGACTAC   1400
T  T  Q  E  E  T  T  T  T  T  T  T  T  E  K  T  T     403

ATCCAAAACTACTACCGAAAAACCAACTACATCAGAATCCGCAACAACGG   1450
 S  K  T  T  T  E  K  P  T  T  S  E  S  A  T  T  E    420
```

```
                ▼         ▼         ▼         ▼
AGACGACAACATCAGAACCTTCAACTACAGAATCAACAACCGTAGATACA   1500
  T  T  T  S  E  P  S  T  T  E  S  T  T  V  D  T     436

                ▼         ▼         ▼         ▼
AGTTCCGCTACGACTGAAGAATCTTCCACAGCCGCGGAGACTACCACTAC   1550
S  S  A  T  T  E  E  S  S  T  A  A  E  T  T  T  T     453

                ▼         ▼         ▼         ▼
ATCTGCTGAAACTTCTGAAACAACAACATCAGAATCTGCAGCTTTTATCA   1600
 S  A  E  T  S  E  T  T  T  S  E  S  A  A  F  I  T    470

                ▼         ▼         ▼         ▼
CCGGAGAATCACCGGAGAATACCGCACTTCAGAGCTCTTCACAAAAGTCT   1650
  G  E  S  P  E  N  T  A  L  Q  S  S  S  Q  K  S     486

                ▼         ▼         ▼         ▼
GAGGAAAATGAATCTTCAGCTGAAAAACCTGGAGCCCGTCGCGATTTTGT   1700
E  E  N  E  S  S  A  E  K  P  G  A  R  R  D  F  V     503

                ▼         ▼         ▼         ▼
TCCAAAGAAACATAAAACTACCGTAAAACCAGCGGAGACCACTTCAGCGG   1750
 P  K  K  H  K  T  T  V  K  P  A  E  T  T  S  A  V    520

                ▼         ▼         ▼         ▼
TAGCAGCTTCAACAACTACCACGGAGCCAATCACCACCACAGAAAAAAGC   1800
  A  A  S  T  T  T  T  E  P  I  T  T  T  E  K  S     536

                ▼         ▼         ▼         ▼
ACTACATTAGAAACTACCCCAATCGAGGCTACTACTTTGAATGAGGTTAC   1850
T  T  L  E  T  T  P  I  E  A  T  T  L  N  E  V  T     553

                ▼         ▼         ▼         ▼
AGGGCCAGCATTCGTAACTGGTGCCCCAGTTGATGGTAATTCGAAATTTC   1900
 G  P  A  F  V  T  G  A  P  V  D  G  N  S  K  F  Q    570

                ▼         ▼         ▼         ▼
AGGCTGATAAATATTACCGATCTCTCTTTTTTTCAGAAACTACTATCAAC   1950
  A  D  K  Y  Y  R  S  L  F  F  S  E  T  T  I  N     586

                ▼         ▼         ▼         ▼
ACTTTAGAATTGCTTAGTAAAATCAATAACACTCAAATTTCGCAGCCAAA   2000
T  L  E  L  L  S  K  I  N  N  T  Q  I  S  Q  P  K     603

                ▼         ▼         ▼         ▼
ACCAACTGACATTTCCAAAACTGATGCACTTCGTCACTTATCAGTGGATT   2050
 P  T  D  I  S  K  T  D  A  L  R  H  L  S  V  D      619

                ▼         ▼         ▼         ▼
AATCGGATCGTTTACCAAAGCTCCAATGGCTCCTACGTGAGTTATACGAA   2100

                ▼         ▼         ▼         ▼
ATTCTTTAATTCTCTAAATTATTTAATTTAATCTTTGATTTTCAGAATAC   2150

                ▼         ▼         ▼         ▼
ATACCACCACCGACGCTGCATTTGTGACTGCAACTGAAGCGTCGTTGAAT   2200
```

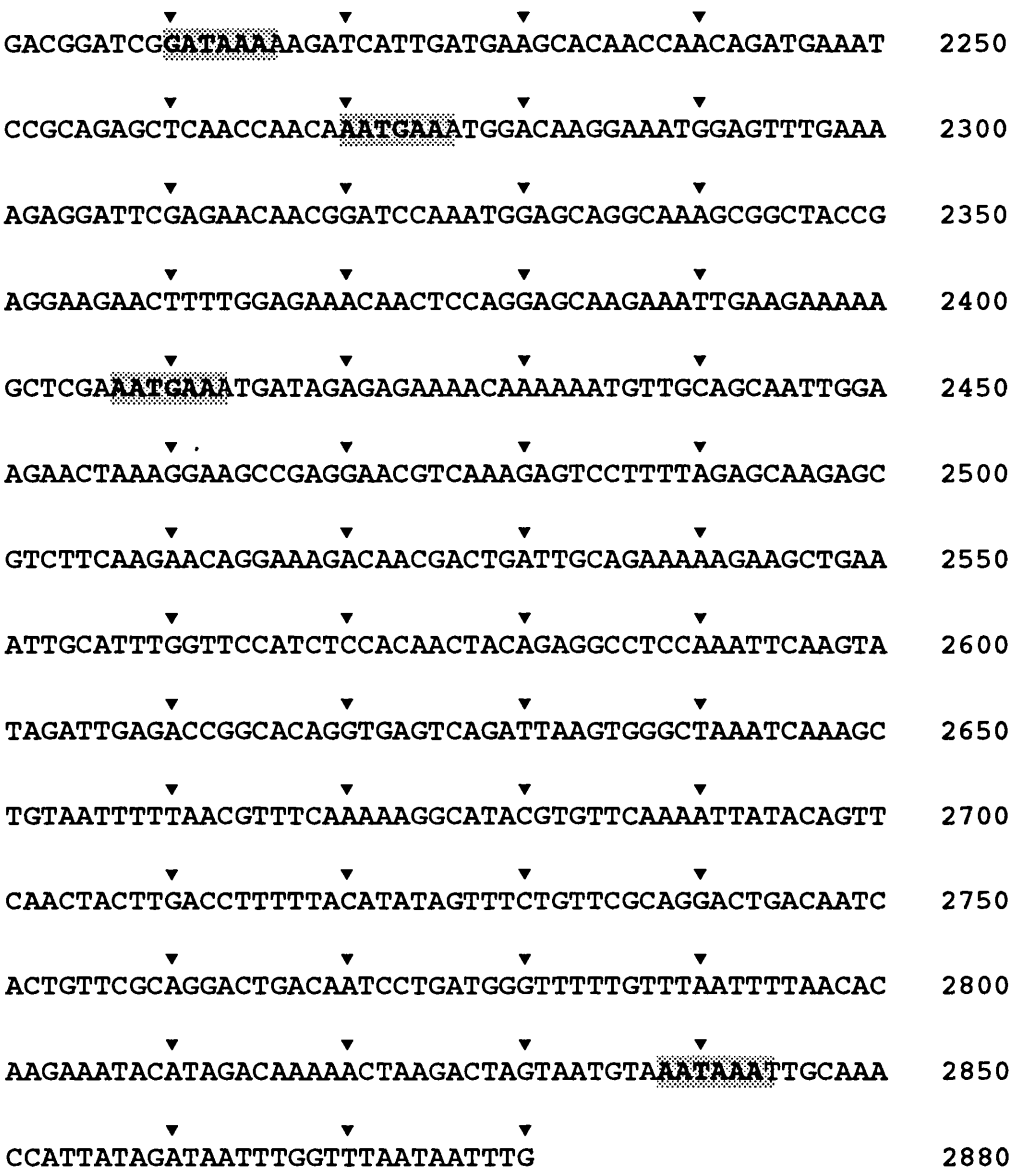


Figure 24. Nucleotide sequence determined from pBlueL and pBlueR.

Some of the highlighted putative sites present in the nucleotide and amino acid sequences are:

-promoter (CAAT and TATA boxes, positions 139 and 152, respectively);

-splicing sites (positions 1069 and 1252); the sequence corresponding to the putative intron is in lower case (as is the amino acid sequence);

-Tc1 insertion site (position 1428);

-polyadenylation signals (positions 2210, 2270, 2407 and 2838);

-potential *N*-glycosylation sites (positions 489 and 595).

communication) can also be found in the 3' end of the sequence (highlighted in Figure 24).

The analysis of the putative protein translated from the largest ORF reveals some features that are suggestive of a functional peptide (Figure 25). The region encompassing the first forty amino acids from the translated protein is mainly hydrophobic, specially the first fifteen amino acids and a region spanning amino acids 20 to 40 is predicted to fold in an alpha-helix. The presence of this N-terminal hydrophobic segment is suggestive of a signal peptide, and considering that there is no other hydrophobic segment of appropriate length that could function as a transmembrane domain, the sequence could code for a secreted protein. The codon usage in the putative protein shows some similarity to the codon usage in abundantly expressed *C. elegans* genes (Table 2). The codon usage in *C. elegans* is highly asymmetric, specially in proline, serine and glycine codons. Although the asymmetry in this sequence is not as pronounced as that reported for other *C. elegans* genes, there is a general similarity in the codon usage for the putative protein.

The amino acid sequence has some unusual features. First, almost 30% of the amino acids are threonine residues. Also present in relatively high amounts are glutamic acid and serine residues. Together, these three amino acids account for more than half of the residues (Table 2). When the amino acid sequence is self-compared using a Diagon plot, the presence of a series of repeats is revealed (Figure 26). Most of the repeats, 34 in total, are arranged in tandem and clustered in the first half of the sequence, near the amino terminus. These six amino acid

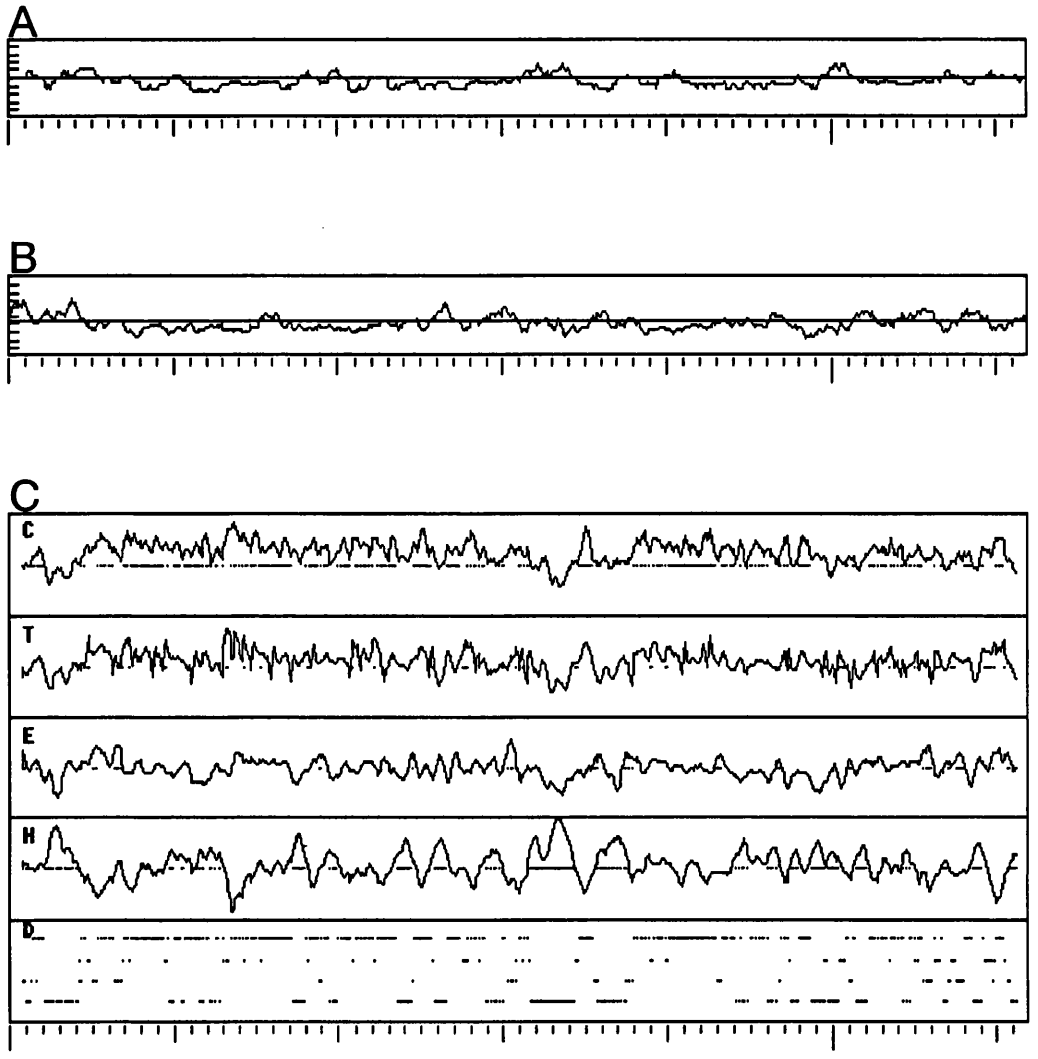


Figure 25. Amino acid sequence analysis.

The graph in panel A corresponds to a charge plot for the amino acid sequence translated from the ORF extending from position 193 to 2049 in the sequence from Figure 24. Panel B shows a hydrophobicity plot using the method described by Kyte and Doolittle (1982). Panel C is a protein secondary structure prediction plot using the Garnier-Osguthorpe-Robson method (Garnier *et al.*, 1978). The method divides protein secondary structures into four classes: coil (C), turn (T), extended or sheet (E) and helix (H). The probability for each residue being found in each of these classes is then calculated and plotted. Below, a strip labelled (D) is divided into four levels, each corresponding to one of the four structure types. For each residue the programme measures which of the four likelihoods is highest and places a single dot at the mid-point of the corresponding level. Each division in the scales beneath the graphs corresponds to 10 amino acids.

Amino Acid	Codon	AA%	codon%	*C. elegans* codon%
Gly	GGG		15.3	0.7
Gly	GGC		15.3	3.3
Gly	GGT		22.9	6.3
Gly	GGA	2.1	46.4	89.6
Glu	GAG		42.4	64.9
Glu	GAA	13.7	57.6	35.1
Asp	GAC		16.6	55.8
Asp	GAT	1.9	83.4	44.2
Val	GTC		10.7	43.8
Val	GTG		10.7	4.9
Val	GTA		39.3	2.1
Val	GTT	4.5	39.3	49.2
Ala	GCC		12.2	44.5
Ala	GCG		15.2	0.9
Ala	GCA		33.2	7.4
Ala	GCT	5.3	39.4	47.2
Arg	CGG		9.1	0.5
Arg	AGG		18.2	1.8
Arg	AGA		18.2	15.6
Arg	CGA		18.2	3.2
Arg	CGT		18.2	46.8
Arg	CGC	1.8	18.2	32.1
Ser	AGC		4.4	6.8
Ser	AGT		8.8	3.6
Ser	TCG		8.8	5.0
Ser	TCC		16.2	35.1
Ser	TCT		25.0	37.0
Ser	TCA	11.0	36.8	12.3
Lys	AAG		40.5	84.2
Lys	AAA	6.8	59.5	15.8
Asn	AAC		46.7	75.6
Asn	AAT	2.4	53.3	24.4
Met	ATG	0.3	100.0	
Ile	ATA		11.9	0.6
Ile	ATT		44.0	35.2
Ile	ATC	4.0	44.0	64.2
Thr	ACG		10.3	1.0
Thr	ACC		23.8	60.2
Thr	ACT		31.9	33.7
Thr	ACA	29.8	34.0	5.1
Trp	TGG	0.2	100.0	

Amino Acid	Codon	AA%	codon%	*C. elegans* codon%
Cys	TGT		0.0	16.9
Cys	TGC	0.0	0.0	83.1
End	TAG		0.0	
End	TGA		0.0	
End	TAA	0.2	100.0	
Tyr	TAC		33.3	78.9
Tyr	TAT	0.5	66.7	21.1
Phe	TTC		28.8	14.9
Phe	TTT	2.3	71.2	85.1
Gln	CAA		37.6	77.9
Gln	CAG	2.6	62.4	22.1
His	CAT		50.0	26.2
His	CAC	1.3	50.0	73.8
Leu	CTC		4.1	35.5
Leu	CTG		8.3	2.0
Leu	CTA		8.3	0.0
Leu	TTA		20.9	0.8
Leu	CTT		20.9	47.4
Leu	TTG	3.9	37.5	14.3
Pro	CCC		2.9	1.6
Pro	CCT		8.8	1.2
Pro	CCG		11.7	2.8
Pro	CCA	5.5	76.6	94.4

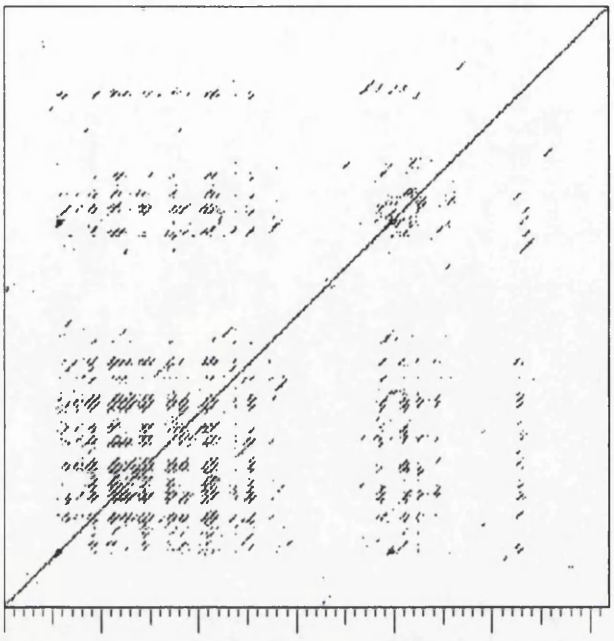
Table 2. Codon usage table.

The codon usage table was compiled for the amino acid sequence translated from the ORF extending from position 193 to 2049 in the sequence from Figure 24. The *C. elegans* codon usage table summarizes data on 3492 codons from the genes for vitellogenins, actin, collagen, major sperm protein, myosin heavy chain and histones (compiled by Emmons, 1988).

AA% - percentage of the particular amino acid;

codon% - codon usage for the protein translated from the ORF;

C. elegans codon% - codon usage for *C. elegans*.



MYSHLLSILISLNQLMRRSFEGVEFIVDFKEKHIILFFRTILHVTVKPRK 50
TTTTTTSTTTTTEKPVITEGSTTNKQPTTSEHESTTEVPTTTEEPTTTTG 100
QHKTEKAITTEEPTTTEESSTTEEVTTTEEPANTGNPPTTENPTTTEQPT 150
STAESTTTALPFTTEQTVTTEEPTTAEKSTATQKPTTTQESVSTEKTSTT 200
KKASTTEEPTTTDEPTTTTESSTTGKATTPELSTTSEETTTTELKITTEG 250
STTTEEPTTTAIFAEASTGIITTDEETTSTTSTTPEITSTKEI*VTESAIT* 300
QTSVSVVESSTPRQLPERWKAIVNKFKHNLEVLKEKKRLLKEKESTSTTG 350
*SDSSE*TTTVVAENIDEVTTTEKEKVVQTTPITTEKSTTQEETTTTTTTTE 400
KTTSKTTTEKPTTSESATTETTTSEPSTTESTTVDTSSATTEESSTAAET 450
TTTSAETSETTTSESAAFITGESPENTALQSSSQKSEENESSAEKPGARR 500
DFVPKKHKTTVKPAETTSAVAASTTTTEPITTTEKSTTLETTPIEATTLN 550
EVTGPAFVTGAPVDGNSKFQADKYYRSLFFSETTINTLELLSKINNTQIS 600
QPKPTDISKTDALRHLSVD 619

Figure 26. Diagon plot and amino acid sequence translated from the largest ORF.

The graph corresponds to a self-comparison of the amino acid sequence using the proportional algorithm of the programme Diagon (Staden, 1982), and contains a representation of all the matches between the pair of sequences (the same sequence being plotted on X and Y axis). In this type of plot, repeats appear as lines parallel to the main diagonal (resulting from perfect identity). The repeats are highlighted in the sequence (bottom). The amino acid sequence translated from the region corresponding to the putative intron is shown in italics. Two other motifs found in some mucin sequences are also underlined. Each division in the scale beneath the graph corresponds to 10 amino acids.

repeats have the consensus sequence TTTEEP (Table 3). However, this consensus is not very well conserved amongst the repeats, but an overall pattern is clearly detectable. The least conserved amino acids are the last two in the repeat unit. This is because the first nucleotide position in these codons is less conserved, with frequencies comparable to the third (wobble) nucleotide position in the other codons (Table 4). Other 12 repeats can be found scattered in the second half of the sequence. Another significant feature in the translated sequence is that the amino acid composition in the region corresponding to where the putative intron is located differs considerably from the surrounding sequences. This region does not have any of the repeats found elsewhere in the sequence, nevertheless is immediately flanked by two of them (Figure 26). Two potential sites for *N*-glycosylation are present in the deduced amino acid sequence, located at positions 489 and 595.

When the current databases are searched using FASTA or TFASTA (Lipman and Pearson, 1985; Pearson and Lipman, 1988), the translated amino acid sequence has no significant homology to any protein. However, the high threonine and serine content is suggestive of a class of proteins named mucins. Mucins and mucin-like glycoproteins have been isolated as the constituents of viscous gel-like matrices covering the surface of the gastrointestinal, urogenital, and tracheobronchial tracts, the oral cavity, and the integument of many species. All mucins contain characteristic serine/threonine-rich domains, which are heavily *O*-glycosylated with a carbohydrate content of up to 80%. When the various mucin (or mucin-like) sequences cloned from different organisms are compared, they show very little or no

AA	1	2	3	4	5	6	Total
T	58.8	82.3	85.3	5.9	0	8.8	40.2
E	0	0	0	64.7	38.2	2.9	17.6
P	2.9	0	2.9	0	0	44.1	8.3
S	14.7	5.9	5.9	0	0	23.5	8.3
K	0	2.9	0	2	20.6	2.9	5.4
A	2.9	2.9	2.9	8.8	0	8.8	4.4
Q	0	0	0	5.9	8.8	2.9	2.9
G	0	0	0	5.9	8.8	0	2.4
V	8.8	0	0	0	2.9	2.9	2.4
I	5.9	2.9	0	0	2.9	0	2.0
N	0	2.9	2.9	0	5.9	0	2.0
L	0	0	0	0	8.8	0	1.5
F	2.9	0	0	0	0	2.9	1.0
H	2.9	0	0	0	2.9	0	1.0
D	0	0	0	2.9	0	0	0.5

Table 3. Amino acid sequence composition of the repeat units.

The table presents the amino acid composition of the repeat units from the translated putative protein. Only those repeats arranged in tandem were used to compile this table. Amino acids not shown were absent from all repeats.
The columns in the table are:
AA - amino acid residue (one letter code);
1 to 6 - percentage of the particular amino acid in the indicated position of the repeat unit;
Total - percentage of the particular amino acid in all positions of the tandem repeats.

N	1	2	3	4	5	6
T	17.6	17.6	26.5	2.9	2.9	41.2
C	5.9	**79.4**	50.0	0	**88.2**	38.2
A	**64.7**	2.9	20.6	**94.1**	5.9	14.7
G	11.8	0	2.9	2.9	2.9	5.9

N	7	8	9	10	11	12
T	5.9	0	38.2	0	0	5.9
C	2.9	**97.1**	8.8	5.9	14.7	8.8
A	**88.2**	2.9	29.4	11.8	**79.4**	47.1
G	2.9	0	23.5	**82.3**	5.9	38.2

N	13	14	15	16	17	18
T	5.9	14.7	11.8	26.5	5.9	0
C	14.7	0	2.9	47.1	**85.3**	2.9
A	29.4	**76.5**	20.6	11.8	8.8	**85.3**
G	50.0	8.8	**64.7**	14.7	0	11.8

Table 4. Nucleotide frequency for the different codon positions in the tandem repeat units.

The frequencies higher than 60% are highlighted. Only those repeats arranged in tandem were used to compile this table.

The columns in the table are:

N- nucleotide;

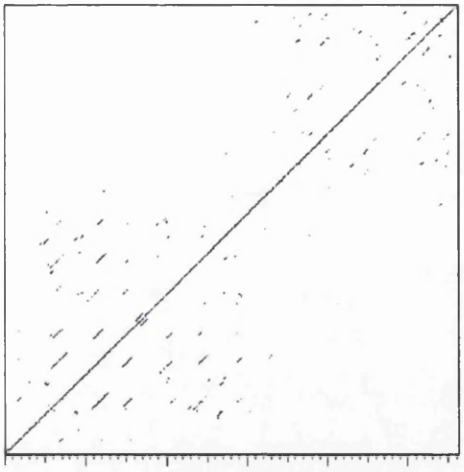
1 to 18 - percentage of a particular nucleotide in the indicated codon position of the repeat units.

conservation at the primary structure level. However, all mucins have in common the presence of a series of repeats which are mainly composed of threonine and serine residues and often arranged in tandem. Figure 27 shows Diagon plots of nine mucin sequences self-compared and the corresponding amino acid sequence with the repeat motifs highlighted. The sequence for rat mucin (Figure 27G) has a very similar repeat motif. There are 59 repeats arranged in tandem, with a consensus sequence TTTPDV. Two other motifs, AVAA and DISK, present in the amino acid sequence translated from the largest ORF are also found in the human and *Xenopus laevis* mucins (Figure 27B, D and H).

3.2.8. Analysis of upstream sequences by *lacZ* translational fusions.

Since the sequence data suggested that the entire coding sequence including regulatory elements were cloned, the potential involvement of sequences upstream of the methionine codon in the expression of the putative protein was analysed by constructing and testing via DNA transformation of *C. elegans*, *lacZ* translational fusions using the pPD21.28 vector. The approach involved cloning the putative regulatory sequences including the first three amino acids in-frame with the *lacZ* gene in pPD21.28.

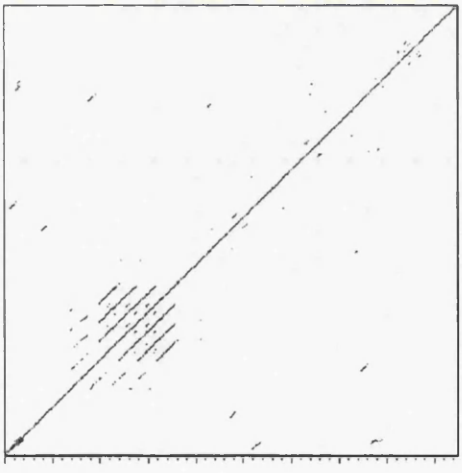
The strategy comprised designing and synthesizing three oligonucleotide primers to be used in PCR reactions to amplify specific segments from pBlueL (Figure 28), which



A

Bovine submaxillary gland mucin-like protein.
Bhargava *et al.* (1990).

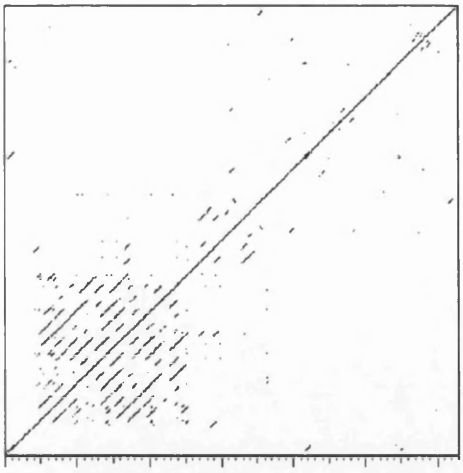
```
MKVLQENSPRHAISGSSHTEATTLIVSNSTSGTGLRPEDNTAVAGGQATG  50
RVTGTTKVIPGTTVAPGSSNTESTTSLGESRTRIGRITGATTGTSKRSSP 100
GSKTGNTGAISGTTVAPGSSNTGATTSLGSGETTQGGIKIVTMGVTTGTT 150
IAPGSSNTKATTPTEVRTTTEVRTATETTTSRHSSDATGSGIQTGITGTG 200
SGTTSSPGGFNAEATTFKEHVRTTETRILSGTTRGRSGTTVIPESSNTGT 250
STGVGRQTSTAVVSGRVTGVSESSSPGTSKEASETTTGPGISTTGSTSKS 300
NRITTSSRIPYPETTVVATGEQETETKTGCTTSLPPPPACYGPLGEKKSP 350
GDIWTANCHKCTCTDAETVDCKLKECPSPPTCKPEERLVKFKDNDTCCEI 400
AYCEPRTCLFNNNDYEVGASFADPKNPCISYSCHNTGFVAVVQDCPKQTW 450
CAEEDRVYDSTKCCYTCKPYCRSSSVNVTVNYNGCKKKVEMARCAGECKK 500
TIKYDYDIFQLKNSCLCCQEENYEYREIDLDCPDGGTIPYRYRHIITCSC 550
LDICQQSMTSTVS                                       563
```



B

Human polymorphic epithelial mucin.
Lancaster *et al.* (1990).

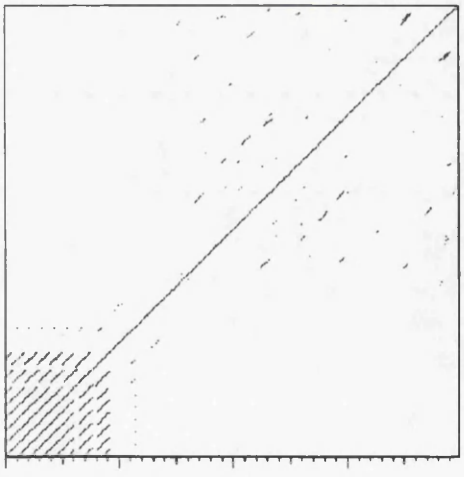
```
MTPGTQSPFFLLLLLTVLTVVTGSGHASSTPGGEKETSATQRSSVPSSTE  50
KNAVSMTSSVLSSHSPGSGSSTTQGQDVTLAPATEPASGSAATWGQDVTS 100
VPVTRPALGSTTPPAHDVTSAPDNKPAPGSTAPPAHGVTSAPDTRPAPGS 150
TAPPAHGVTSAPDNRPALGSTAPPVHNVTSASGSASGSASTLVHNGTSAR 200
ATTTPASKSTPFSIPSHHSDTPTTLASHSTKTDASSTHHSTVPPLTSSNH 250
STSPQLSTGVSFFFLSFHISNLQFNSSLEDPSTDYYQELQRDISEMFLQI 300
YKQGGFLGLSNIKFRPGSVVVQLTLAFREGTINVHDVETQFNQYKTEAAS 350
RYNLTISDVSVSDVPFPFSAQSGAGVPGWGIALLVLVCVLVALAIVYLIA 400
LAVCQCRRKNYGQLDIFPARDTYHPMSEYPTYHTHGRYVPPSSTDRSPYE 450
KVSAGNGGSSLSYTNPAVAATSANL                           475
```



C

Mouse (*Mus musculus*) mammary epithelium mucin.
Spicer *et al.* (1991).

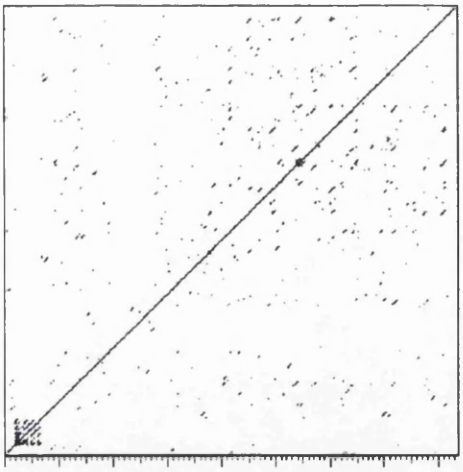
```
MTPGIRAPFFLLLLLASLKGFLALPSEENSVTSSQDTSSSLASTTTPVHS  50
SNSDPATRPPGDSTSSPVQSSTSSPATRAPEDSTSTAVLSGTSSPATTAP 100
VNSASSPVAHGDTSSPATSPLKDSNSSPVVHSGTSSAATTAPVDSTSSPV 150
VHGGTSSPATSPPGDSTSSPDHSSTSSPATRAPEDSTSTAVLSGTSSPAT 200
TAPVDSTSSPVAHDDTSSPATSLSEDSASSPVAHGGTSSPATSPLRDSTS 250
SPVHSSASIQNIKTTSDLASTPDHNGTSVTTTSSALGSATSPDHSGTSTT 300
TNSSESVLATTPVYSSMPFSTTKVTSGSAIIPDHNGSSVLPTSSVLGSAT 350
SLVYNTSAIATTPVSNGTQPSVPSQYPVSPTMATTSSHSTIASSSYYSTV 400
PFSTFSSNSSPQLSVGVSFFFLFFYIQNHPFNSSLEDPSSNYYQELKRNI 450
SGLFLQIFNGDFLGISSIKFRSGSVVVESTVVFREGTFSASDVKSQLIQH 500
KKEADSYNLTISEVKVNEMQFPPSAQSRPGVPGWGIALLVLVCILVALAI 550
VYFLALAVCQCRRKSYGQLDIFPTQDTYHPMSEYPTYHTHGRYVPPGSTK 600
RQPYEEVSAGNGSSSLSYTNPAVVTTSANL                      630
```



D

Xenopus laevis integumentary mucin FIM-B.1 (partial sequence).
Probst *et al.* (1990).

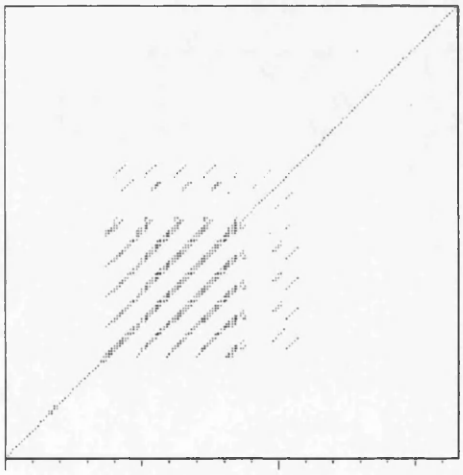
```
PAPSETTGESTPAPSETTGESTPAPSETTGESTPAPSETTGESTPAPSET  50
TGESTPAPSETTVPSGESTPAPSETTVPSVPSGESTPAPSETTELRIIPP 100
EVSTVAVPVTTGQITPAVTTEHSTEEILTLPPPVVGPVLPAKPTVDISKY 150
TNTTTTKSTVPTTTIPPKATCCGSSGESVQAGHMWQTGCDVCTCNGTSGK 200
TQCAPRQCEKEIICKSDERRVLRKPGKSCCGYCEPLTCKHNGTEYKLGAT 250
FIDKSNPCITYRCDASGLTVNVKSCPNEQVCSKSERTYDSDGCCFSCDTS 300
CKPVPATVGIQGEYDYQNEKTNCSANIIMAKCSGQCQHKLTYDTIDNKVV 350
TKCRCCKADRVEPRKAHLVCDNGKKKIYKYKHITSCKCTSCTAYNIRL   398
```



E

Rat (*Rattus sp.*) mucin-like peptide (partial sequence). Xu *et al.* (1991).

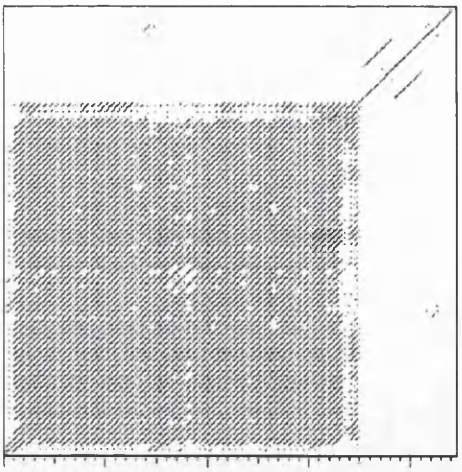
```
VNCSVDCQLQVFNWSCPSTPSTPPPSTPTTPTSSQTTTPSTPSTTSSKST  50
PSTPQSTSSKSTPSTPPKTTLPGCLDFDPPRQVNETWWLCNCTMAICKYD 100
NVVEIVELECNPPPMPTCSNGLKPVRVPDPDGCCWHWECDCYCTGWGDPH 150
FVTFDGLYYSYQGNCTYVLVEEITPTVDNFGVYIDNYHCDANDKVSCPRT 200
LIVHHETQEVLIKTVHMMPIEVEVQVNKQLVALPYKKYGLEVYQSGINFV 250
VDIPRLGAQVSYNGLSFSIRLPYHLFGNNTKGQCGTCTNNTADDCILPSG 300
EIISNCEVAADEWLVNDPSKPHCPHKGLTTKRPAITTPGPFPENCTVSPV 350
CQLIMDSLFSQCHPFVPPKHYYEACLFDSCFVAGSGMECASVQAYAALCA 400
QEGVCIDWRNHTQGACAVTCPAHRQYQACGPSEEPTCQSSSPKNSTLLVE 450
GCFCPEGTTKFAPGYDVCVKICGCVGPDNVPREFGEHFEFDCKDCVCLEG 500
GSGIVCQPKKCARGNLTTCEEDGTYLVVEADPDDKCCNTTSCKCDPKRCK 550
AERPSCLLGFEVKSEHVPGKCCPVYSCVPKGVCVHENAEYQPGSPVYSNK 600
CQDCVCTDSMDNSTQLNVISCTHVPCNISCSSGFELVEVPGECCKKCQQT 650
HCIIKRPEQQYIILKPGEIQKNPNDRCTFFSCMKINNQLISSVSNITCPD 700
FDPSDCVPGSITYMPNGCCKTCIHNPNNTVPCSAIPVMKEISYNGCAKNI 750
SMNFCAGSCGTFAMYSAQAQDLDHGCSCCREERTSVRMVSLDCPDGSKLS 800
HSYTHIESCLCQGTVCELPQAQQSRTRRSSPRLLGRK               837
```



F

Trypanosoma cruzi mucin. Reyes *et al.* (unpublished, EMBL database accession number L20809).

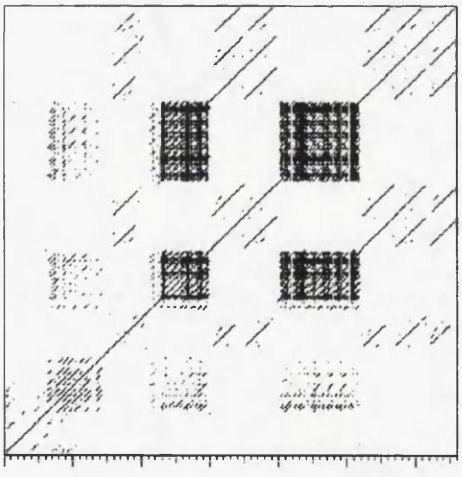
```
MNTLTMMTCRLLYVLLVLALCCCPSVCTTESGQEETTTTTTTTKSPTTTT  50
TTTTTKPPTTTTTTTTTKPPTTTTTTTTTTKPPTTTTTTTTTTAPEAPSITTTET 100
PNTTTTRAPSSIRRIDGSLGSSAWACARCFSPHPRWRTPLWAEEVCAGCA 150
CEHSATGKCVLLPWSDV                                   167
```



G

Rat (*Rattus norvegicus*) intestinal mucin (partial sequence). Gum *et al.* (1991).

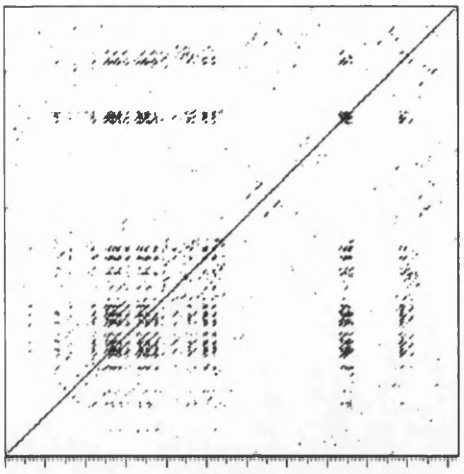
```
GTTTVDFTTTAEFTTTPDVTTTPDVTTTPGVTTTAGVTTTPDVTTTAGVT    50
TTPDVTTTADVTTTPDVTTTADVTTTAGITTTPDVTTTAGVTTTPDVTTT   100
PDVTTTPDVTTTPDVTTTPDVTTTPDTTTTPDTTTTPQVTTTVDVTTTVD   150
VTTTAEVTTTTEVTTSPDATTTTDVTTTPEATTTDGVTTTPDVTTTSDVT   200
TTADVTTTASVTTTPDVTTTPDVTTTPDVTTTAGVTTTPDATTTPDATTT   250
PQVTTTADVTTTAGVTTTAGVTTTPDVTTTPDVTTTPDVTTTASVTTTAD   300
VTTTPEVTTTPEVTTTPAVTTTPRVTTTPAVTTTRVITTMPVITTTPVMT   350
TTPVNCLNGGYWSGAMCVCPNGFSGDRCQNRVPVVDCQNGGTWDGLKCQC   400
TGLFYGPRCEEVMESVEIKPTVSASVEVSVTVTSQEYSNELQDRNST      447
```



H

Xenopus laevis integumentary mucin FIM-C.1 (partial sequence). Hauser and Hoffmann (1992).

```
APTTAAAVAATGKDTTAAAEGSAAAEKTAAAGEVSAPPTAAVAATGEDAT    50
TAAATAAAETTAAAGEAPTTTTAPATTAAGKAPTTAAATAPTTAAAGAPT   100
TATGKAPATAAAPVPTTAASKAPTTAAAATHSTAAAAAPTTAASAAKSKE   150
RSTSSSSEEEHCHVKPSKREMCGSKGITKKQCKKKNCCFDPKGHGGIHCF   200
HRKPKGHSHEEHTTTTTKAPTTIQIATTTTTTPTTTTTTTTTKATPTTTTTTK   250
ATPTTTTTTTKATTTTTTTTPTTTTTTTTTKATTTPTTTTTTTTPTTTTTKATTTTT   300
TTTTSGECKMEPSKREDCGYSGITESQCRTKGCCFDSSIPQTKWCFYTLSQ   350
VADCKVEPSQRVDCGFRGITADQCRQKNCCFDSSISGTKWCFYSTSQVAA   400
TKTTTTPTTTTTPTTTTTTTKATTTTPTTTTTTTPTTTTTTTTTTTKATTTTP   450
TTTTPTTTTTKATTTTTPTTTTTTTTPTTTTTTKATTTTPTTTTTTPTTTTTKA   500
TTTTTPTTTTTTTTTTTTTKATTTTTSGECKMEPSKRADCGYPGITESQCRSKG   550
CCFDSSIPQTKWCFYSLPQVADCKVAPSSRVDCGFGGITADQCRQRNCCF   600
DSSISGTKWCFYSTSQGNAMCSGPPTKRRDCGYPGISSSVCINRGCCWDN   650
SVMNVPWCFYRT                                          662
```



Canine tracheobronchial mucin (partial sequence).
Shankar *et al.* (1992).

```
NPGTGAGCSRMSILKIHAREIFDSRGIPTVEVDLYTTKGLFRAAVPSGAS   50
TGIYEALELVTMTRTPTGKVSQKAVETSIKTIAPALISKNVNVVGAKTRL  100
TTTDAGTWMDQRTRASGTTTYPPQAGPQTPTPLSTPASSSGPQPRFYPSS  150
VLSAVSETTPTTTTLRPYPSSVLICCSLTTPSTAPGEEVYNGGIRYSSLS  200
IPSSAPPQTADDPPQCLTPTPTPLSTPSIIRPQPEALTLPLCLSAVSETT  250
PTPISTTTTVTPTPTPTGTQTPTTTPITTTTVTPTPTPTGTQTPTTVLIT  300
TTTTMPTPTPTSQKLRGSDTPVLVIALVTPTPTPTGTQTPTTTPITTTTT  350
VTPTPTPTGTQTPTTIRITTTTTMTPTPTTTSTKVLVTPYTTTTGVTPTT  400
HRHTRHQTTTTHTTYTTVTAIATPTGTRPTTYLHTTTTVTQTTHRQTTRP  450
QPTTPSPTTTTVDPTQHPSPPPLGDPNPNTHRDTDPNHDTTTTTTVTPTP  500
TPTRTRPQPRYYTTTTTMTPTPTPPTPHTDHHHDPLPPPPPQVRHHRTPR  550
HCHELTSGSPLTHYPNTPTTNLLEKNRSELINTPTHTMIISSAAMATPPS  600
LVHSRCRSGHSSKAQRKSRASGCGLMMLGVESGVGVGSTHAGGALVRSRC  650
RFTRHSRAEEGLGPEAVVSMMLGVESGVGVGVRHCGGRLHSAGGGAEDDP  700
GRVPRVGALTDSTQHTSRKAIKCKAGVPNEELTTYIACASALQSETCAAD  750
MIATATIHTTTSRSIRKPKTISELTTYACASTAASVGTVVQLDMIATGGL  800
RDRGQMAGAMPAGDSRGSPGTELEFRPNGLQAPTPTPISTTTTVTPTPTP  850
TGTQTPTTTPSPPPPGDPTKGKTVYKIIGSLRTTFNLPAWLGYITICKCL  900
MQSNRKTKTAETINVKTLGSLMMDVTPINSCRVTPPFPSGKPCADSRLMN  950
PTAETGSVIGAGCSCCVLTTPTTHRVIRCTTAPTPTPISTTTTVTPTQHP 1000
PAHRPTTSPITTTTTVTQPDTTHTDPTTYIPTTRCRIEKPVERSSRCGKT 1050
MGCGPTLLSAGVGGVGCRGCCALSGCVSEVAGRGTGQNQTGPCPIIEALA 1100
KYNQLLRLKRLGSKAPFAGRNFRNPRIN                        1128
```

Figure 27. Diagon plots and corresponding amino acid sequence from different mucins.

The amino acid sequence from nine different mucins was self-compared using the proportional algorithm of Diagon (Staden, 1982). The amino acid sequences are shown under the graphs, with the repeat motifs highlighted. Two other motifs, AVAA and DISK (sequences B, D and H), that are also found in the *C. elegans* putative protein are doubly underlined. Each division in the scale beneath the graphs corresponds to 10 amino acids.

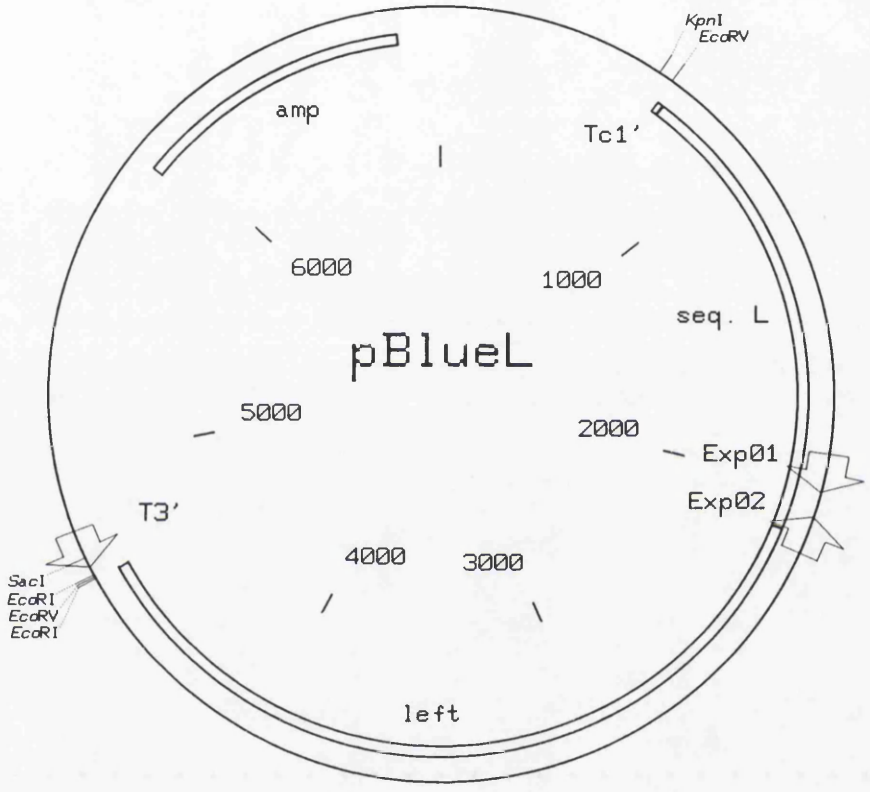


Figure 28. Strategy for cloning sequences upstream of the methionine codon in the putative protein.

A map of the plasmid pBlueL is shown together with the location of oligonucleotide primers T3', Exp01 and Exp02 (indicated by arrows) used to amplify two fragments from pBlueL to be cloned into the pPD21.28 expression vector. There are two regions marked in the insert: **seq.L**, corresponding to parts where the sequence has been determined, and **left**, not sequenced. The location of the ampicillin resistance gene from pBluescript is also indicated.

contains the amino terminus of the putative protein. The three primers designed were:

1) Exp01, containing a *Bam*HI site and hybridizing to the region where the putative start codon is located;

2) Exp02, containing a *Pst*I site and hybridizing to the sense DNA strand approximately 200 bp upstream of Exp01;

3) T3', hybridizing to the region in pBluescript where the T3 promoter is located, approximately 2.5 kb upstream of Exp01.

Figure 29 shows the sequence and the hybridizing sites for these three primers. They would allow the amplification of two DNA segments, from Exp01 to Exp02, and from Exp01 to T3'. Both PCR products would have a *Pst*I site at one end[13] and a *Bam*HI site located in the Exp01 primer. This primer includes the first three codons from the putative protein, which could then be cloned into the *Bam*HI site present in the polylinker of pPD21.28 while maintaining the reading frame respectively to *lacZ* (Figure 30).

ssDNA from pBlueL was used as the target for PCR amplification in two separate reactions, using the Exp01/Exp02 and Exp01/T3' primers. The PCR products (Figure 31A) were purified by successive extractions with phenol, phenol-chloroform and chloroform, followed by ethanol precipitation. The purified DNA was digested with *Bam*HI and *Pst*I and ligated to *Bam*HI/*Pst*I digested pPD21.28. Competent XL1-Blue cells were then independently transformed with both constructs and plated on L-agar plates containing ampicillin, X-Gal and IPTG. Most of the transformant colonies were pale

---

[13] located either in Exp02 or in the polylinker of pBluescript.

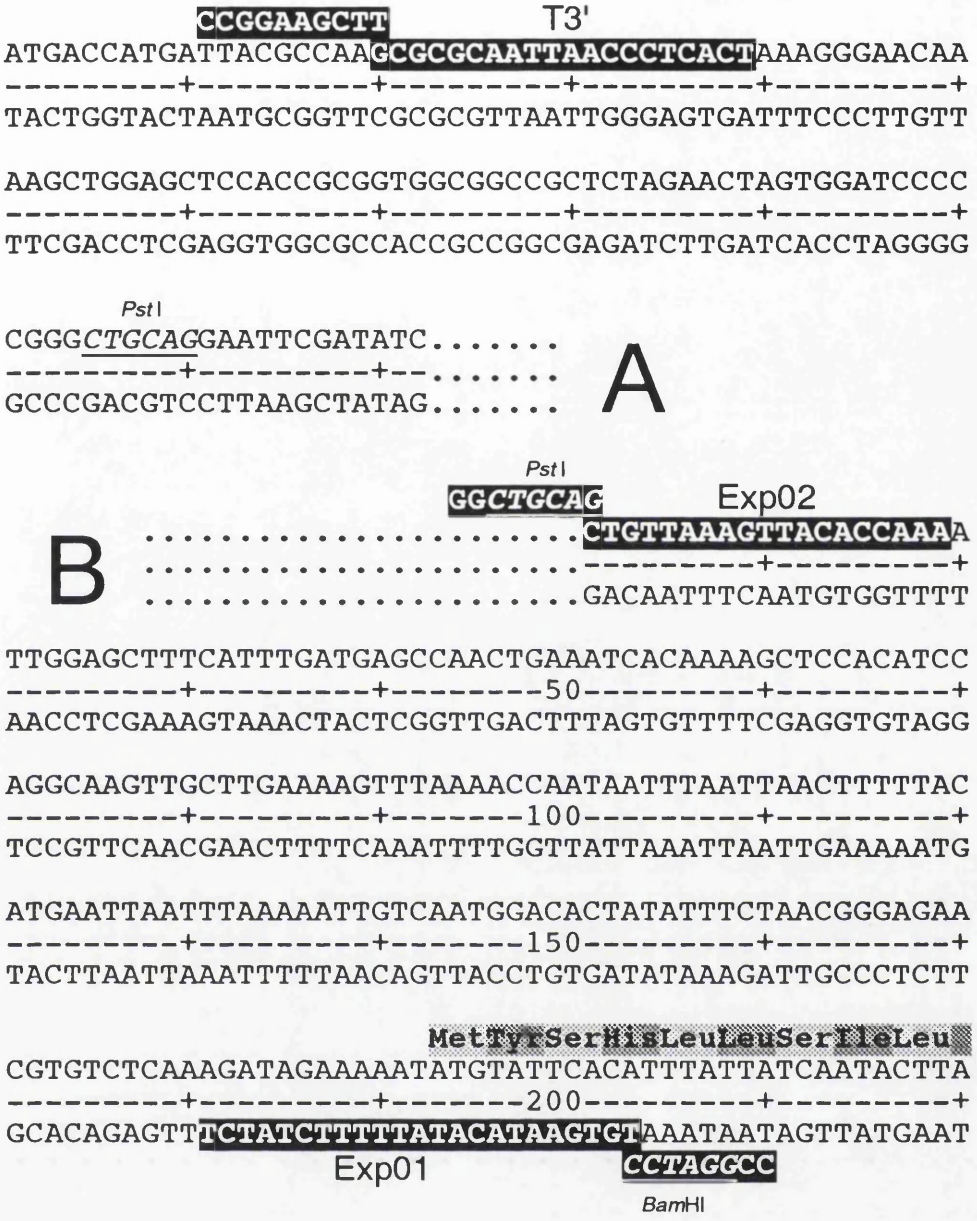


Figure 29. Sequence and location of the oligonucleotide primers T3', Exp01 and Exp02.

Partial sequence data from pBlueL are shown. The sequence in panel A corresponds to pBluescript, up to the *Eco*RV site in the vector/insert boundary. The dotted lines represent parts of the cloned insert that were not sequenced. The sequence in panel B corresponds to the sequenced part of pBlueL close to the putative start codon (the numbering is identical to that of Figure 24). The first nine amino acids are indicated. The primers are outlined in black. The *Bam*HI and *Pst*I sites in Exp01 and Exp02, respectively, are shown underlined in italics. Also highlighted is the *Pst*I site downstream from the T3' primer which was used for cloning the larger PCR product amplified using the Exp01/T3' primers.

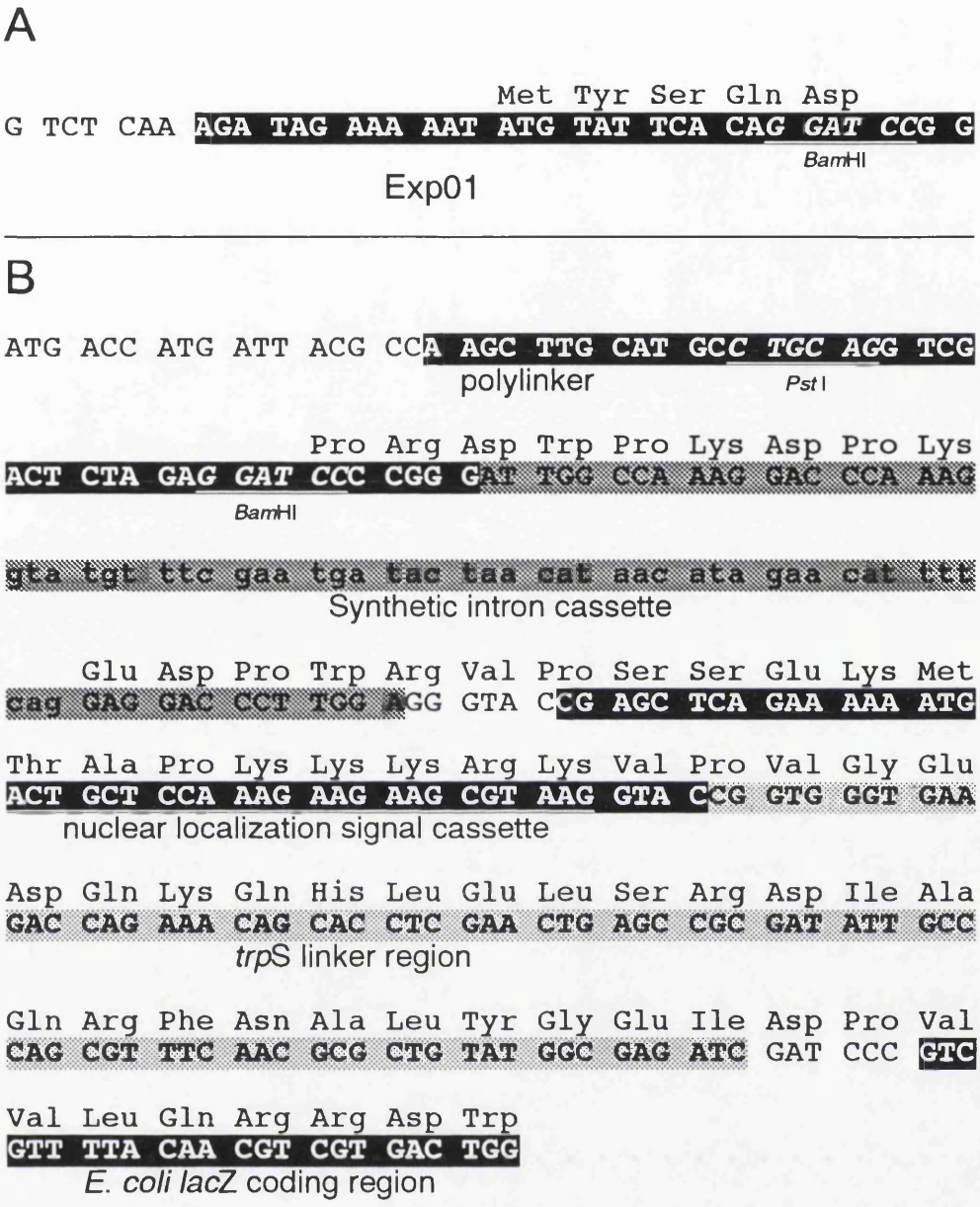


Figure 30. pPD21.28 translational fusion constructs.

Panel A corresponds to the last 38 bases of the predicted sequence for a PCR product amplified using Exp02 or T3' and Exp01 primers. The first five amino acids are indicated, up to the *Bam*HI site. Panel B corresponds to the partial sequence from pPD21.28, from the polylinker up to the first eight amino acids from *lac*Z. The translated amino acid sequence, starting immediately after the *Bam*HI site, is also indicated. The intron sequence in the synthetic intron cassette is shown in lower case. The eight codons corresponding to the nuclear localization signal are underlined. All relevant individual components (cassettes) of the vector are indicated under the respective highlighted sequences.

blue, together with a few blue and white colonies. When plasmid DNA was prepared, digested with restriction enzymes and analysed by agarose gel electrophoresis, only the pale blue colonies contained the desired constructs. Figure 31B shows the DNA prepared from some recombinant colonies digested with *Bam*HI and *Pst*I, and electrophoresed in an agarose gel. The construct containing the 2.5 kb fragment was named pPRO-B and the other construct, containing the 200 bp fragment, was named pPRO-S.

Both constructs, pPRO-B and pPRO-S, were independently injected into *C. elegans* N2 animals. Ten hermaphrodites were injected with each plasmid (together with pRF4, the plasmid containing the *rol*-6 gene), and more than one hundred F3 animals exhibiting roller phenotype were then stained with X-Gal. Both plasmids failed to produce animals with identifiable nuclear staining.

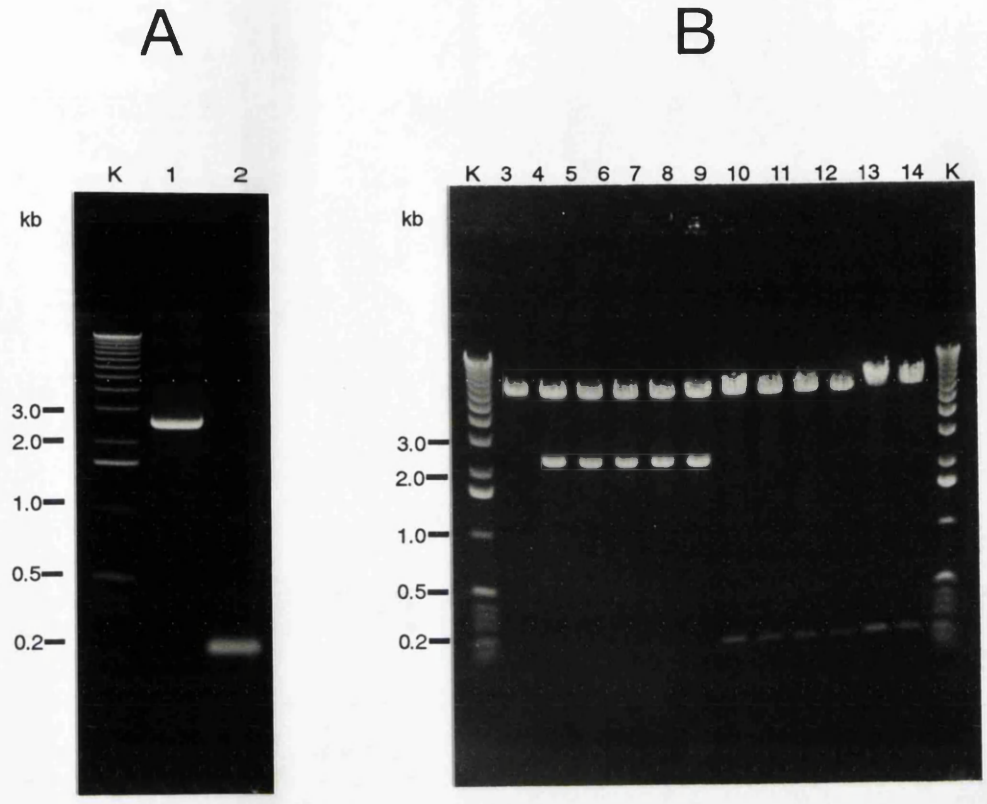


Figure 31. PCR amplification of sequences upstream of the methionine codon in the putative protein and cloning into pPD21.28.

Panel A shows an agarose gel electrophoresis of the PCR products obtained when using the T3'/Exp01 (lane 1) and Exp02/Exp01 (lane 2) primer combinations to amplify sequences from the pBlueL plasmid. 2 ng of pBlueL ssDNA was used in the PCR reactions. Annealing was performed at 50°C (30 seconds), followed by primer extension (72°C for 1 minute) and denaturation (94°C for 30 seconds), during 30 cycles. Samples corresponding to 1/20 of the PCR reaction were withdrawn and electrophoresed in a 0.6% agarose gel. The PCR products were digested with *Bam*HI and *Pst*I and cloned into *Bam*HI/*Pst*I digested pPD21.28. After transformation of XL1-Blue cells, DNA from several colonies was prepared, digested with *Bam*HI and *Pst*I and analysed by electrophoresis in a 0.7% agarose gel (panel B).

K) 1 kb ladder; 1) PCR using T3'/Exp01 primers; 2) PCR using Exp02/Exp01 primers; 3) DNA from pPD21.28; 4-8) DNA prepared from recombinant colonies carrying the cloned PCR product amplified using T3'/Exp01 primers; 9-14) DNA prepared from recombinant colonies carrying the cloned PCR product using Exp02/Exp01.

4. DISCUSSION.

4.1. The collagenolytic activity in *C. elegans* is associated with the cuticle material.

The cuticle of *C. elegans* is composed mainly of distinct collagen molecules (Cox *et al.*, 1981b; Ouazana and Herbage, 1981; Edgar *et al.*, 1982). During the life cycle four moults occur, when a new cuticle is synthesized and the old cuticle is shed. Until now there is no information available regarding the involvement of collagenases or other proteinases in the moulting process. It is not clear if the release of the old cuticle is a simple mechanical process, driven by the synthesis of a new cuticle by the hypodermal cells, or if some kind of matrix degradation takes place.

The results from the assays using substrate gels indicated that there was a collagenolytic activity in *C. elegans* and that it was present associated only with the cuticle. No activity could be detected in the supernatant from sonicated worms. From the two methods used for treating cuticle preparations, the collagenase activity could only be detected in the extracts prepared using Calcium to dissociate it from the cuticle. Despite the fact that incubation with ß-ME is very efficient in solubilizing the cuticle proteins (Cox *et al.*, 1981c), it could also be detrimental to the collagenase activity.

The substrate gels were performed using collagen copolymerised in the polyacrylamide gel instead of the more common substrate gelatin, ensuring a high degree of specificity in the assays. Only collagenases have the ability to cleave intact,

fibrillar collagen. Once the higher order structure is destroyed, however, many proteases, including other metalloproteases, can degrade the denatured collagen/gelatin substrates (Matrisian, 1992). The assays performed using casein-containing substrate gels indicated that the enzyme present in *C. elegans* had no activity against this substrate. Together, these results suggested that the collagenolytic activity found associated with the cuticle material in *C. elegans* could represent a true collagenase.

Considering that the collagenolytic activity detected in *C. elegans* is associated exclusively with the cuticle material, and since most of the proteins in the cuticle are collagens, some kind of regulation of the enzyme activity must exist so that substrate degradation would occur in a controllable way. Such a mechanism could be the post-translational regulation controlling the activity of metalloproteases which has been described in higher eukaryotes and *Paracentrotus lividus* (sea urchin). In this mechanism, known as the cysteine switch hypothesis (Nagase *et al.*, 1990; Springman *et al.*, 1990), the cysteine residue present in the activation site is coordinated to the active-site Zinc atom and the disruption of this interaction by physical or chemical means is the first step in the activation of these enzymes. Recently, Salowe *et al.* (1992) demonstrated the existence of a Zinc-Sulphur coordination in the pro-enzyme of Stromelysin, the first experimental evidence that the mechanism proposed by the cysteine switch hypothesis is possible.

If a mechanism similar to the cysteine switch could be in operation in *C. elegans*, it might be predicted that there would be conservation of amino acids not only in the active centre of the enzyme, but also in an activation site similar to that

described for other metalloproteases.

Based on the assumption that the putative metalloprotease activity detected in *C. elegans* could have the same conservation at the primary structure level as that found in equivalent enzymes from other organisms, two oligonucleotides, specific for these conserved sites, were designed. However, due to the presence of amino acids with highly degenerate codons in these sites, the oligonucleotides had to incorporate inosine in some positions, reducing their specificities. When used in PCR experiments, these oligonucleotides were able to amplify the expected product when the plasmid containing the cloned sea urchin collagenase was used as target DNA. However, no product was amplified when human or sea urchin genomic DNA was used. Also, these primers failed to detect specific fragments containing homologous sequences when used to probe Southern blots containing human and sea urchin DNA. Presumably, the reduced specificity prevented the hybridization to the corresponding sequences in the target DNAs.

An alternative approach for cloning the putative *C. elegans* collagenase would be screening a cDNA expression library using antibodies. This experimental approach has been successfully used for cloning the sea urchin collagenase, using polyclonal antibodies raised against the purified enzyme to screen a λgt11 library (Lepage and Gache, 1990). Presumably, being polyclonal these antibodies could recognize some common epitopes in the putative *C. elegans* enzyme and could also be used for detecting the *C. elegans* enzyme.

It remainsto be determined, however, if the detected

collagenase activity associated with the cuticle of *C. elegans* plays any role in the moulting process. Nevertheless, the idea that some enzymatic degradation of the old cuticle takes place and in some way helps the process of shedding seems to be a reasonable assumption. The molecular cloning of the putative collagenase from *C. elegans* would be the first step towards elucidating its functions and whether it could be involved in the moulting process.

4.2. The molecular cloning of a sequence potentially encoding a mucin-like protein.

Transposon tagging was the procedure used for cloning the *dpy*-6 gene from *C. elegans*. This approach has been successfully used to clone a number of *C. elegans* genes including *unc*-22 (Moerman *et al.*, 1986), *dpy*-13 (von Mende *et al.*, 1988), *lin*-12 (Greenwald, 1985), *fem*-3 (Rosenquist and Kimble, 1988), *sqt*-1 (Kramer *et al.*, 1988), *mec*-3 (Way and Chalfie, 1988) and *unc*-7 (Starich *et al.*, 1993) among others.

The strain used, DR1013, is a mutator strain where, following a spontaneous transposition event, a Tc1 transposon is disrupting the *dpy*-6 gene activity. After replacing the DR1013 chromosomal background with N2 chromosomes, a single Tc1 insertion was found associated with the dumpy phenotype. This Tc1 insertion was located in a 9.5 kb *Eco*RI fragment, which physically maps to a region in chromosome X of *C. elegans* corresponding to the genetic map position of the *dpy*-6 gene. Also, P. Albert (personal communication) had previously

mapped, by complementation crossings, the Tc1 transposon tagged gene to the *dpy*-6 locus. Both results, the physical mapping and the complementation analysis, suggested that the tagged gene was *dpy*-6.

However, the cosmid containing the wild-type 7.9 kb *Eco*RI fragment equivalent to the Tc1 tagged allele failed to rescue the dumpy phenotype of CB14 [*dpy*-6(e14)] and WT-C [*dpy*-6(m445)] animals when used in transformation experiments. One possible explanation for this result could be that the cosmids used in the microinjections contained small deletions or other rearrangements. The occurrence of deletions in cosmid preparations is not uncommon and this has been shown to be the case in failed attempts to rescue the gene *par*-2 (D. Levitan, C. Mello and D. Stinchcomb, personal communication[1]). Also, H. Browning and S. Strome (personal communication[2]) suggested that deletions or rearrangements in cosmids could be the reason for failed attempts to rescue the gene *spe*-11 using cosmids spanning the region where the gene was located. One viable alternative to avoid the problems with deletions in cosmids would be by using YACs in the microinjections. A disadvantage in using YACs for phenotype rescue is that, due to their large size, several genes could be included in the cloned fragment, making it difficult to establish the precise identity of the gene being rescued. Also, attempts to rescue the phenotype could be performed by first isolating the corresponding wild-type gene from a genomic library and using

---

[1] Worm Breeder's Gazette, 11(1):20, September/1989.

[2] Worm Breeder's Gazette, 12(3):41, June/1992.

it in the microinjections. Another distinct possibility that cannot be ruled out based on the results is that the tagged gene is not *dpy*-6, but a closely-linked gene.

The analysis of the cloned sequence suggested that an entire coding sequence had been cloned. An ORF of 1857 bp was found, with the Tc1 transposon inserted approximately in the middle. This putative coding sequence is flanked by sequences similar to regulatory elements described for other *C. elegans* genes.

Located 41 bp upstream of the putative start codon there is the sequence TATATT, which is very similar to the TATA box consensus sequence. The location of this sequence is comparable to that of the TATA box found in *C. elegans* vitellogenin genes. In these genes, a homologous sequence is found 39 to 42 bp upstream of the start codon, depending on the particular *vit* gene (Spieth *et al.*, 1985).

A sequence identical to VPE1[3], also from vitellogenin genes, is found 54 bp upstream of the putative start codon. This promoter element, which is similar to the mammalian CCAAT box, has been shown to have a regulatory function in the expression of the vitellogenin *vit*-2 gene (MacMorris *et al.*, 1992). A 247 bp segment of *vit*-2 5'-flanking DNA is sufficient for correctly regulated, high-level expression of this gene. It has been demonstrated that the VPE1 immediately upstream from the TATA box is the most conserved of these elements, in terms of sequence and location, in all of the *vit* genes (Spieth *et al.*, 1985). They are located at 57 bp (*vit*-1), 58 bp (*vit*-2), 59 bp

---

[3] vit promoter element, which has the consensus sequence TGTCAAT.

(*vit*-4) and 58 bp (*vit*-5) from the start codon. A 2-bp change in this promoter element (changing it to TG**GC**CAAT), by site-directed mutagenesis, completely eliminated promoter function.

However, no homology to the VPE2 consensus sequence (CTGATAA) was found. This second promoter element found in vitellogenin genes is identical to the binding site for the mammalian erythrocyte-specific transcriptional activator GATA-1 (Orkin, 1990) and has also been found upstream of the TATA box of the major sperm protein from *C. elegans* (Klass *et al.*, 1988), in a region required for intestine-specific expression of the *C. elegans* gut esterase gene *ges*-1 (Aamodt *et al.*, 1991), and just upstream of the *C. elegans tra*-2 gene required for correct sex determination (Okkema and Kimble, 1991).

In vitellogenin genes, both promoters elements, VPE1 and VPE2, function as activating sequences, suggesting that they are binding sites for activator proteins. Mutations in some of these elements always result in variable reductions in expression levels. None of the VPEs act like repressor-binding sites, since mutations in these elements do not cause an increase in gene expression. Also, they do not appear to be involved in tissue, stage or sex-specific regulation of expression.

Together, the presence of sequences homologous to the TATA box and VPE1 promoter element of *C. elegans*, located in similar places upstream of the methionine codon, is suggestive that they could represent functional promoter elements in the putative gene.

However, when *lacZ* translational fusions including these sequences were microinjected into N2 animals, no

expression of the fusion protein could be detected by staining with X-Gal. Two different constructs were tested. The first included about 200 bp upstream of the putative start codon and the second included more than 2.5 kb of DNA. In the case of the smaller construct it could be argued that it did not include enough upstream sequences to drive expression of the *lacZ* fusion protein. However, considering the promoter structure described for other *C. elegans* genes (Xue *et al.*, 1992; Aamodt *et al.*, 1991) and the size of the larger construct, it would be expected that all the required elements of a functional promoter should be included in it. Another possible explanation for the failure in detecting nuclei staining, other than that the upstream sequences do not contain a functional promoter, is that since they were cloned from PCR amplified products they could include mutations impairing the promoter activity. This could have been the case since DNA from a single construct for each plasmid was used in the microinjections. The translational fusions were constructed including only the first three amino acids of the putative protein. This would avoid the inclusion of the putative signal sequence in the fusion proteins and the potential for being secreted. However, secretion cannot be excluded, which would cause a lack of detectable X-Gal staining.

The nucleotide sequence downstream of the putative stop codon of the ORF contains a sequence identical to the classical polyadenylation signal present in other eukaryotes. Apart from this sequence, other putative polyadenylation signals are also present. Recently, T. Blumenthal, O. White and

C. Fields (personal communication[4]) examined cDNA sequences from about 1300 different *C. elegans* genes and proposed that almost 44% of these genes could use polyadenylation signals alternative to the consensus. The most frequently found sequences were AATGAA (8.3%), TATAAA (7.4%), CATAAA (5%) and GATAAA (4.8%). Other variations in the consensus signal, containing a 1-base mismatch, were also found. For some of the genes analysed no putative polyadenylation signal was found.

The ORF is interrupted by what seems to be an intron, 186 bp in length. The suggestion that this sequence could be an intron comes from two structural features. First, the putative intron is flanked by sequences similar to the consensus splice acceptor and donor sites described in *C. elegans*. These sites are situated in such a way that excision of the putative intron would maintain the reading frame. Second, the amino acid composition deduced from the putative intron is considerably different from that of the rest of the ORF. As will be discussed later, the region of the putative protein adjacent to the intron has a peculiar primary structure, with the presence of a series of amino acid repeats with a characteristic composition, which is completely absent from the sequence translated from the putative intron.

The amino acid sequence translated from the putative coding sequence is similar to that of mucin-like proteins. Mucin glycoproteins are expressed by a wide variety of epithelial tissues where they serve as a protective layer covering the plasma membrane. They have also recently been identified in the cell membrane of *Trypanosoma cruzi*, where they are sialylated and

---

[4] Worm Breeder's Gazette, 13(1):62-63, October/1993.

participate in the process of mammalian cell invasion (Schenkman *et al.*, 1993; Schenkman and Eichinger, 1993).

Mucin and mucin-like proteins are a subclass of glycoproteins, which are characterized by two main features: first, a high percentage of their weight consists of oligosaccharides, *O*-linked to serine or threonine residues in the polypeptide backbone, and second, the protein backbone contains a number of repeating sequences, including virtually all the *O*-linked oligosaccharide attachment sites (Strous and Dekker, 1992; Rose, 1992; Devine and McKenzie, 1992). Some mucin polypeptides are also *N*-glycosylated (Hilkens and Buijs, 1988; Amerongen *et al.*, 1987; Dekker *et al.*, 1989), most probably in the N- and C-terminus, outside the repetitive region. Asparagine residues in glycoproteins are only *N*-glycosylated if they are part of the consensus sequence NXS or NXT, where X can be any amino acid except proline and aspartic acid (Marshall, 1972).

A further subdivision of mucin-like glycoproteins is of potential importance for their function: there are secretory and membrane-bound types.

The secretory mucin-like glycoproteins are characterized by the presence of cysteine-rich domains. These domains are thought to be responsible for oligomerization via interchain disulfide bridges (Dekker and Strous, 1990). This process is supposed to be important for the gaining of viscous gel-forming properties of secreted mucins. A mainly linear assembly is thought to occur end-to-end (Dekker *et al.*, 1991; Strous and Dekker, 1992), but polymerization via non-mucin "link" proteins has also been suggested (Allen, 1983).

The membrane-bound type of mucins have a hydrophobic stretch of amino acid residues anchoring the long filamentous molecules in the plasma membrane. This class of mucins probably lack intermolecular disulfide bridges and their functions are less well understood, but they could have a role in cell adhesion and recognition. For pancreatic and mammary mucins, it has been shown that no cysteine residues are present in the polypeptide extending from the cell surface; any cysteine residues are present in the region of the protein spanning the lipid bilayer (Gendler *et al.*, 1990; Ligtenberg *et al.*, 1990; Lan *et al.*, 1990).

Recently, Lasky *et al.* (1992) have identified a novel, atypical human endothelial mucin-like peptide which, despite the characteristic high serine and threonine content, contains neither the repetitive motifs nor the cysteine-rich domains and hydrophobic transmembrane segment. They have proposed that this molecule could be incorporated into the glycocalyx as a peripheral membrane protein by association with a transmembrane component or attachment could be achieved by oligomerization of the C-terminal amphipathic helical regions and insertion of these regions into the membranes so that the polar regions interact with each other to form an oligomer and the apolar faces of the helices interact with the lipid bilayer.

The sequence analysis of the putative protein from *C. elegans* is very suggestive of a mucin-like glycoprotein. It has the characteristic tandem repeats found in the mucin sequences. These repeats, which in most of the mucins analysed are rich in threonine and serine residues interspersed by proline residues, are arranged in tandem near the amino-terminus of the protein

and have been shown to vary considerably between mucins of different species and also between different mucins of the same species.

The repeats present in the deduced amino acid sequence from *C. elegans* have some similarity with those from rat intestinal mucin (Gum *et al.*, 1991). Like the rat mucin sequence, it has a tandem repeat unit which is six residues in length, threonine being the residue most frequently found in the first three positions of the repeat unit. Similarly, some mucin sequences also have this stretch of three or more threonine residues in the repeat unit. The last three residues in the repeat unit are less conserved, but glutamic acid in positions 4 and 5, and proline in the last position are the most frequent residues. The last two residues in the repeat units are the least conserved of all. This is due to lack of conservation in the first nucleotide position of this codon. The nucleotide frequencies in these codon positions are comparable to those in the third position of the first four codons in the repeat units. Gum *et al.* (1991) found a similar situation in one codon of the rat mucin repeat. They proposed two possible explanations for this phenomenon that could also be applied for the *C. elegans* putative gene. First, point mutations could have occurred in the primordial gene at a time when only a few tandem repeats were present. Then, tandem repeats with different mutations would be propagated by gene duplications, resulting in the differences seen in the modern gene. Second, there is selective pressure keeping a combination of certain amino acids in this positions (E and K in position 5; P and S in position 6).

The region in the putative protein corresponding to the

repeat units would constitute the sites for *O*-glycosylation. Also, *N*-glycosylation could occur at two sites located at positions 489 and 595 in the amino acid sequence. It is worth noting that these two possible sites for *N*-glycosylation are located outside the tandem repeats.

Two other motifs found in the deduced amino acid sequence from *C. elegans* are also present in other mucin sequences. The motif AVAA is found in the human polymorphic epithelial mucin and in the *Xenopus laevis* integumentary mucin FIM-C.1. The motif DISK is found in the *Xenopus laevis* integumentary mucin FIM-B.1. However, these motifs are not found in any other mucin sequences and their significance has not been determined.

There are no cysteine-rich domains present in the deduced amino acid sequence from *C. elegans*. In fact, the sequence is devoid of any cysteine residues. This lack of cysteine residues indicates that mechanisms involving oligomerization via disulfide bonds could not occur. Also, there is neither any long stretch of hydrophobic amino acids that could function as a transmembrane domain for anchoring the protein nor any amphipathic helix closer to the carboxi-terminus that could serve similar function.

The putative protein from *C. elegans* could resemble the "surface-associated" proteins and glycoproteins, which are easily solubilized and can be dissociated from the cuticle without disulfide reduction. Its mucin-like structure and a possible association with the dumpy phenotype is suggestive that it could form part of the *C. elegans* cuticle structure.

Zuckerman *et al.* (1979) have partially characterized

the cuticle surface carbohydrates from *C. briggsae* and *C. elegans* and found the presence of galactose, glucose, mannose, and *N*-acetylglucosamine on the exposed surface. They have found also an apparent absence of digestible glycoproteins, as suggested by the results of protease digestion studies. However, they have suggested that this could be due to the presence of the surface coat.

More recently, glycoproteins have been demonstrated to form part of the *C. elegans* cuticle and that alterations in their expression could lead to abnormalities in the cuticle.

Using a monoclonal antibody, Hemmer *et al.* (1991) have detected a stage-specific antigen on the surface of the first larval stage (L1) of *C. elegans* that was sensitive to digestion by pronase and *O*-glycanase, suggesting that this antigen is an *O*-linked glycoprotein. In two *C. elegans* mutants, *srf*-2 and *srf*-3, the antigen is not detected on the surface of animals from any larval stage (Politz *et al.*, 1990). The *srf*-2 and *srf*-3 phenotypes appear to be loss of a stage-specific antigenic determinant. Another mutation, *srf*-(yj43), caused the antigen to be expressed on the surface of the first through the fourth (L4) larval stages. Politz *et al.* (1990) have suggested that this phenotype could result either from failure to down-regulate L1-specific antigen expression at later stages or persistence of the synthesized antigen. The antigen-defective phenotype of *srf*-3 was epistatic to the heterochronic mutant phenotype of *srf*-(yj43) in immunofluorescence tests of the *srf*-3/*srf*-(yj43) double mutant, suggesting that *srf*-(yj43) causes incorrect regulation of a pathway of antigen formation that requires wild-type *srf*-3 activity. None of these mutations caused any striking alterations

in overall morphology. However, another set of *srf* genes (*srf*-4, *srf*-8 and *sfr*-9), identified by ectopic surface binding of the lectins wheat germ agglutinin and soybean agglutinin (Link *et al.*, 1992), showed pleiotropic phenotypes which included an abnormal male copulatory bursae, defective egg-laying, uncoordinated movement, and gonadal defects. They suggested that merely an altered cuticle could not account for all the abnormalities observed, and that a possible alteration in hypodermal function could be involved. These three loci could affect a single pathway involved in cell-cell interaction or cell positional information, perhaps through a defect in glycosylation of extracellular matrix components.

Blaxter (1993), analysing the molecular components of the surface of *C. elegans* by radioiodination has identified a single, non-collagenous, hydrophobic protein which is not glycosylated and is a heterodimer of 6.5 kDa and 12 kDa subunits. Surface iodination of *rol*-6[5] animals did not reveal any differences from the N2 wild-type strain. However, iodination of *dpy*-3[6] mutants resulted in very low levels of incorporation, the heterodimer being faintly detectable. Also, as the heterodimer was absent from the previously mentioned *srf*-4, *srf*-8 and *srf*-9 mutants, one might speculate that an alteration in the expression of glycoproteins in the *C. elegans* cuticle could have some effect on the expression, processing or assembling of other cuticular proteins.

Another example of the presence of glycoproteins in the

---

[5] the *rol*-6 gene has been determined to be a cuticular collagen (Kramer *et al.*, 1990).

[6] the nature of the mutation in the *dpy*-3 gene (Brenner, 1974) has not been determined.

cuticle of nematodes has been found in *Toxocara canis*. Page *et al.* (1992) have identified a labile antigenic surface coat covering the epicuticle of infective larvae. From observations by electron-microscopy, this coat seems to be similar to that described by Zuckerman *et al.* (1979) for *C. elegans*. The major antigen associated with this coat is a glycoprotein of 120 kDa (TES-120), which is known to be extensively glycosylated (Meghji and Maizels, 1986). The mucin-like nature of this coat component was further suggested by a high incorporation of ^{14}C-labelled serine and threonine into TES-120 (Page and Maizels, 1992). Recently, D. Gems and R. Maizels (personal communication[7]) cloned the gene encoding TES-120. The encoded product can be divided into three different parts: a 16 amino acid residue N-terminal hydrophobic signal peptide; an 86 amino acid residue mucin-like serine/threonine-rich domain; and a C-terminal cysteine-rich domain. The mucin-like domain contains 72.1% serine and threonine residues, largely contained in a number of repeats of the consensus sequence STSSSSA.

Considering the structural similarities of the putative protein from *C. elegans* with mucin glycoproteins, it is conceivable that it could represent a cuticular mucin-like protein. Such a protein could either form part of the cuticle structure or be located on the surface of hypodermal cells, interacting with proteins in the basal zone of the cuticle. One point that needs further elucidation is whether this putative protein is encoded by the *dpy*-6 gene. However, it is feasible that the altered expression of a cuticular glycoprotein could as a

---

[7] Worm Breeder's Gazette, 13(1):83, October/1993.

result cause a disruption in the organization of the cuticle and consequently induce a dumpy phenotype. Also, despite the fact that the identity of the Tc1 tagged gene has not been definitely established to be *dpy*-6, the results indicate that the isolated Tc1 insertion, with the consequent disruption of the putative mucin-like gene, is associated with the dumpy phenotype. Nevertheless, one could speculate that the putative *C. elegans* mucin-like protein is part of the surface coat overlaying the epicuticle, in which case a direct correlation with the dumpy phenotype would be much less apparent.

One conceivable way of clarifying the possible role of the putative mucin-like protein from *C. elegans* in the cuticle structure would be by using antibodies as probes for identifying its location in purified cuticles or in intact animals.

5. BIBLIOGRAPHICAL REFERENCES.

Aamodt, E.J., Chung, M.A. and McGhee, J.D. (1991) Spatial control of gut-specific gene expression during *Caenorhabditis elegans* development. *Science*, 252:579-582.

Allen, A. (1983) Mucus - a protective secretion of complexity. *Trends in Biochemical Sciences*, 8:169-173.

Alting-Mees, M.A. and Short, J.M. (1989) pBluescript II: gene mapping vectors. *Nucleic Acids Research*, 17:9494.

Amerongen, A.V.N., Oderkerk, C.H., Roukema, P.A., Wolf, J.H., Lisman, J.J.W. and Vliegenthart, J.F.G. (1987) Primary structure of *O*- and *N*-glycosylic carbohydrate chains derived from murine submandibular mucin. *Carbohydrate Research*, 164:43-48.

Benton, W.D. and Davis, R.W. (1977) Screening lambda gt recombinant clones by hybridization to single plaques "in situ". *Science*, 196:180-182.

Bhargava, A.K., Woitach, J.T., Davidson, E.A. and Bhavanandan, V.P. (1990) Cloning and cDNA sequence of a bovine submaxillary gland mucin-like protein containing two distinct domains. *Proceedings of the National Academy of Sciences of USA*, 87:6798-6802.

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) Cuticle surface proteins of wild-type and mutant *Caenorhabditis elegans*. *Journal of Biological Chemistry*, 268:6600-6609.

Boedtker, H., Fuller, F. and Tate, V. (1983) The structure of collagen genes. *International Review on Connective Tissue Research*, 9:1-63.

Bornstein, P. and Sage, H. (1980) Structurally distinct collagen types. *Annual Review of Biochemistry*, 49:957-1003.

Breathnach, R., Matrisian, L.M., Gesnel, M.-C., Staub, A. and Leroy, P. (1987) Sequences coding for part of oncogene-induced transin are highly conserved in a related rat gene. *Nucleic Acids Research*, 15:1139-1151.

Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics*, 77:71-94.

Bullock, W.O., Fernandez, J.M. and Short, J.M. (1987) XL1-Blue: a high efficiency plasmid transforming *rec*A *Escherichia coli* strain with beta-galactosidase selection. *Biotechniques*, 5:376-379.

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

Clarke, L. and Carbon, J. (1976) A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome. *Cell*, 9:91-99.

Collier, I.E., Wilhelm, S.M., Eizen, A.Z., Marmer, B.L.,

Grant, G.A., Seltzer, J.L., Kronberger, A., He, C., Bauer, E.A. and Goldberg, G.I. (1988) H-ras oncogene-transformed human bronchial epithelial cells (TBE-1) secrete a single metalloproteinase capable of degrading basement membrane collagen. *Journal of Biological Chemistry*, 263:6579-6587.

Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. and Chambon, P. (1980) Promoter sequences of eukaryotic protein-coding genes. *Science*, 209:1406-1414.

Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of USA*, 83:7821-7825.

Coulson, A., Waterston, R.H., Kiff, J., Sulston, J.E. and Kohara, Y. (1988) Genome linking with yeast artificial chromosomes. *Nature*, 335:184-186.

Cox, G.N., Laufer, J.S., Kush, M. and Edgar, R.S. (1980) Genetic and phenotypic characterization of roller mutants of *Caenorhabditis elegans*. *Genetics*, 95:317-339.

Cox, G.N., Kusch, M., DeNevi, K. and Edgar, R.S. (1981a) Temporal regulation of cuticle synthesis during development of *Caenorhabditis elegans*. *Developmental Biology*, 84:277-285.

Cox, G.N., Kusch, M. and Edgar, R.S. (1981b) The cuticle of *Caenorhabditis elegans*. Its isolation and partial

characterization. *Journal of Cell Biology*, 90:7-17.

Cox, G.N., Silvija, S. and Edgar, R.S. (1981c) The cuticle of *Caenorhabditis elegans*. II. Stage-specific changes in ultrastructural and protein composition during postembryonic development. *Developmental Biology*, 86:456-470.

Cox, G.N., Kramer, J.M. and Hirsh, D. (1984) Number and organization of collagen genes in *Caenorhabditis elegans*. *Molecular and Cellular Biology*, 4:2389-2395.

Cox, G.N., Carr, S., Kramer, J.M. and Hirsh, D. (1985) Genetic mapping of *Caenorhabditis elegans* collagen genes using DNA polymorphisms as phenotypic markers. *Genetics*, 109:513-528.

Cox, G.N. and Hirsh, D. (1985) Stage-specific patterns of collagen gene expression during development of *Caenorhabditis elegans*. *Molecular and Cellular Biology*, 5:363-372.

Cox, G.N., Fields, C., Kramer, J.M., Rosenzweig, B. and Hirsh, D. (1989) Sequence comparisons of developmentally regulated collagen genes of *Caenorhabditis elegans*. *Gene*, 76:331-344.

Cross, S.H. and Little, P.F.R. (1986) A cosmid vector for systematic chromosome walking. *Gene*, 49:9-22.

Dekker, J., van Beurden-Lamers, W.M.O. and Strous, G.J. (1989) Biosynthesis of gastric mucus glycoprotein of the

rat. *Journal of Biological Chemistry*, 264:10431-10437.

Dekker, J. and Strous, G.J. (1990) Covalent oligomerization of rat gastric mucin occurs in the rough endoplasmic reticulum, is *N*-glycosylation-dependent, and precedes initial *O*-glycosylation. *Journal of Biological Chemistry*, 265:18116-18122.

Dekker, J., van der Ende, A., Aelmans, P.H. and Strous, G.J. (1991) Rat gastric mucin is synthesized and secreted exclusively as filamentous oligomers. *Biochemical Journal*, 279:251-256.

Denhardt, E.M. (1966) A membrane filter technique for the detection of complementary DNA. *Biochemical and Biophysical Research Communications*, 23:641-646.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research*, 12:387-395.

Devine, P.L. and McKenzie, F.C. (1992) Mucins: structure, function and associations with malignancy. *Bioessays*, 14:619-625.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. and Mattick, J.S. (1991) Touchdown PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research*, 19:4008.

Edgar, R.S., Cox, G.N., Kusch, M. and Politz, J.C. (1982) The cuticle of *Caenorhabditis elegans*. *Journal of Nematology*,

14:248-258.

Edwards, M.K. and Wood, W.B. (1983) Location of specific messenger RNAs in *Caenorhabditis elegans* by cytological hybridization. *Developmental Biology*, 97:375-390.

Emmons, S.W. and Yesner, L. (1984) High-frequency excision of transposable element Tc1 in the nematode *Caenorhabditis elegans* is limited to somatic cells. *Cell*, 36:599-605.

Emmons, S.W. (1988) The genome. In *The nematode Caenorhabditis elegans*. W.B. Wood (ed.), Cold Spring Harbour (NY): Cold Spring Harbour Laboratory Press.

Feinberg, A.P. and Vogelstein, B. (1983) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochemistry*, 132:6-13.

Feinberg, A.P. and Vogelstein, B. (1984) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity (addendum). *Analytical Biochemistry*, 137:266-267.

Fessler, J.H. and Fessler, L.I. (1978) Biosynthesis of procollagen. *Annual Review of Biochemistry*, 47:129-162.

Fetterer, R.H. and Rhoads, M.L. (1990) Tyrosine-derived cross-linking amino acids in the sheath of *Haemonchus contortus* infective larvae. *Journal of Parasitology*, 76:619-624.

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

Fini, M.E. Plucinska, I.M. Mayer, A.S. Gross, R.H. and Brinckerhoff, C.E. (1987) A gene for rabbit synovial cell collagenase: member of a family of metalloproteinases that degrade the connective tissue matrix. *Biochemistry*, 26:6156-6165.

Fire, A. (1986) Integrative transformation of *C. elegans*. *EMBO Journal*, 5:2673-2680.

Fire, A., Harrison, S.A. and Dixon, D. (1990) A modular set of *lacZ* fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene*, 93:189-198.

Frischauf, A.M., Lehrach, H., Poustka, A. and Murray, N. (1983) Lambda replacement vectors carrying polylinker sequences. *Journal of Molecular Biology*, 170:827-842.

Fujimoto, D. and Kanaya, S. (1973) Cuticlin: a non-collagen structural protein from *Ascaris* cuticle. Archives of Biochemistry and Biophysics, 157:1-6.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120:97-120.

Gendler, S.J., Lancaster, C.A., Taylor, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E. and Wilson, D. (1990) Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *Journal of Biological Chemistry*, 265:15286-15293.

Goldberg, G.I., Wilhelm, S.M., Kronberger, A., Bauer, E.A., Grant, G.A. and Eizen, A.Z. (1986) Human fibroblast collagenase: complete primary structure and homology to an oncogene transformation-induced rat protein. *Journal of Biological Chemistry*, 261:6600-6605.

Greenwald, I. (1985) *lin*-12, a nematode homeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. *Cell*, 43:583-590.

Grunstein, M. and Hogness, D.S. (1975) Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. *Proceedings of the National Academy of Sciences of USA*, 72:3961-3965.

Gum, J.R.,Jr., Hicks, J.W., Lagace, R.E., Byrd, J.C., Toribara, N.W., Siddiki, B., Fearney, F.J., Lamport, D.T.A. and Kim, Y.S., (1991) Molecular cloning of rat intestinal mucin: lack of conservation between mammalian species. *Journal of Biological Chemistry*, 266:22733-22738.

Harris, E.D. and Vater, C.A. (1982) Vertebrate collagenases. In *Methods in Enzymology (Structural and Contractile Proteins)*. L.W. Cunningham and D.W. Frederiksen (eds.), New York: Academic Press, p.423-452.

Hauser, F. and Hoffmann, W. (1992) P-domains as shuffled cysteine-rich modules in integumentary mucin C.1 (FIM-C.1) from *Xenopus laevis*: polydispersity and genetic polymorphism. *Journal of Biological Chemistry*, 267:24620-24624.

Hemmer, R.M., Donkin, S.G., Chin, K.J., Grenache, D.J., Bhatt, H. and Politz, S.M. (1991) Altered expression of an L1-specific , *O*-linked cuticle surface glycoprotein in mutants of *Caenorhabditis elegans*. *Journal of Cell Biology*, 115:1237-1247.

Heussen, C. and Dowdle, E.B. (1980) Electrophoretic analysis of plasminogen activators in polyacrylamide gels containing sodium dodecyl sulphate and copolymerised substrates. *Analytical Biochemistry*, 102:196-202.

Higgins, D.G. and Sharp, P.M. (1988) Clustal: a package for performing multiple sequence alignments on a microcomputer. *Gene*, 73:237-244.

Higgins, D.G. and Sharp, P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *CABIOS*, 5:151-153.

Hilkens, J. and Buijs, F. (1988) Biosynthesis of MAM-6, an epithelial sialomucin. Evidence for involvement of a rare proteolytic cleavage step in the endoplasmic reticulum. *Journal of Biological Chemistry*, 263:4215-4222.

Hodgkin, J., Horvitz, H.R. and Brenner, S, (1979) Nondisjunction mutants of the nematode *Caenorhabditis elegans*. *Genetics*, 91:67-94.

Hodgkin, J. (1983) X chromosome dosage and gene expression in *Caenorhabditis elegans*: two unusual dumpy genes. *Molecular and General Genetics*, 192:452-458.

Hodgkin, J. (1988) Sexual dimorphism and sex determination. In *The nematode Caenorhabditis elegans*. W.B. Wood (ed.), Cold Spring Harbour (NY): Cold Spring Harbour Laboratory Press.

Hope, I.A. (1991) "Promoter trapping" in *Caenorhabditis elegans*. *Development*, 113:399-408.

Innis, M.A., Gelfand, D.H., Sninsky, J.J. and White, T.J. (1990) *PCR protocols, a guide to methods and applications*. San Diego: Academic Press, (USA).

Ish-Horowicz, D. and Burke, J.F. (1981) Rapid and efficient cosmid cloning. *Nucleic Acids Research*, 9:2989-2998.

Johnstone, I.L., Shafi, Y. and Barry, J.D. (1992) Molecular analysis of mutations in the *Caenorhabditis elegans* collagen gene *dpy-7*. *EMBO Journal*, 11:3857-3863.

Kingston, I.B., Wainwright, S.M. and Cooper, D. (1989) Comparison of collagen gene sequences in *Ascaris suum* and *Caenorhabditis elegans*. *Molecular and Biochemical Parasitology*, 37:137-146.

Kramer, J.M., Cox, G.N. and Hirsh, D. (1982) Comparisons of the complete sequences of two collagen genes from *Caenorhabditis elegans*. *Cell*, 30:599-606.

Kramer, J.M., Cox, G.N. and Hirsh, D. (1985) Expression of the *Caenorhabditis elegans* collagen genes *col-1* and *col-2* is developmentally regulated. *Journal of Biological Chemistry*, 260:1945-1951.

Kramer, J.M., Johnson, J.J., Edgar, R.S., Basch, C. and Roberts, S. (1988) The *sqt*-1 gene of *C. elegans* encodes a collagen critical for organismal morphogenesis. *Cell*, 55:555-565.

Kramer, J.M., French, R.P., Park, E.C. and Jonhson, J.J. (1990) The *Caenorhabditis elegans rol*-6 gene, which interacts with the *sqt*-1 collagen gene to determine organismal morphology, encodes a collagen. *Molecular and Cellular Biology*, 10:2081-2089.

Krebs, M.P., Raj Bhandary, U.L. and Khorana, H.G. (1990) Nucleotide sequence of ISH11, a new *Halobacterium halobium* insertion element isolated from the plasmid pGRB1. *Nucleic Acids Research*, 18:6699.

Kretz, P.L. (1989) Effect of lambda packaging extract *mrc* restriction activity on DNA cloning. *Nucleic Acids Research*, 17:5409.

Kush, M. and Edgar, R.S. (1986) Genetic studies of unusual loci that affect body shape of the nematode *Caenorhabditis elegans* and may code for cuticle structural proteins. *Genetics*, 113:621-639.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105-132.

Lackey, A., James, E.R., Sakanari, J.A., Resnick, S.D., Brown, M., Bianco, A.E. and McKerrow, J.H. (1989) Extracellular proteases of *Onchocerca*. *Experimental*

Parasitology, 68:176-185.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227:680-685.

Lan, M.S., Batra, S.K., Qi, W.N., Metzgar, R.S. and Hollingsworth, M.A. (1990) Cloning and sequencing of a human pancreatic tumor mucin cDNA. *Journal of Biological Chemistry*, 265:15294-15299.

Lancaster, C.A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J. and Gendler, S.J. (1990) Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochemical and Biophysical Research Communications*, 173:1019-1029.

Laski, L.A., Singer, M.S., Dowbenko, D., Imai, Y., Henzel, W.J., Grimley, C., Fennie, C., Gillett, N., Watson, S.R. and Rosen, S.D. (1992) An endothelial ligand for L-selectin is a novel mucin-like molecule. *Cell*, 69:927-938.

Lepage, T. and Gache, C. (1990) Early expression of a collagenase-like hatching enzyme gene in the sea urchin embryo. *EMBO Journal*, 9:3003-3012.

Levy, A.D., Yang, J. and Kramer, J.M. (1993) Molecular and genetic analyses of the *Caenorhabditis elegans dpy*-2 and *dpy*-10 collagen genes: a variety of molecular alterations affect organismal morphology. *Molecular Biology of the Cell*, 4:803-817.

Ligtenberg, M.J.L., Vos, H.L., Gennissen, A.M.C. and Hilkens, J. (1990) Episialin, a carcinoma-associated mucin, is generated by a polymorphic gene encoding splice variants with alternative amino termini. *Journal of Biological Chemistry*, 265:5573-5578.

Link, C.D., Silverman, M.A., Breen, M., Watt, K.E. and Dames, S.A. (1992) Characterization of *Caenorhabditis elegans* lectin-biding mutants. *Genetics*, 131:867-881.

Lipman, D.J. and Pearson, W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, 227:1435-1441.

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.

Marshall, R.D. (1972) Glycoproteins. *Annual Review of Biochemistry*, 41:673-702.

Martin, F.H., Castro, M.M., Aboulela, F. and Tinoco, I. (1986) Base-pairing involving deoxyinosine: implications for probe design. *Nucleic Acids Research*, 13:8927-8938.

Matrisian, L.M., Glaichenhaus, N., Gesnel, M.-C. and Breathnach, R. (1985) Epidermal growth factor and oncogenes induce transcription of the same cellular mRNA in rat fibroblasts. *EMBO Journal*, 4:1435-1440.

Matrisian, L.M. (1990) Metalloproteinases and their inhibitors in

matrix remodelling. *Trends in Genetics*, 6:121-125.

Matrisian, L.M. (1992) The matrix-degrading metalloproteinases. *BioEssays*, 14:455-463.

Meghji, M. and Maizels, R.M. (1986) Biochemical properties of larval excretory-secretory (ES) glycoproteins of the parasitic nematode *Toxocara canis*. *Molecular and Biochemical Parasitology*, 18:155-170.

McKerrow, J.H., Pino-Heiss, S., Lindquist, R. and Werb, Z. (1985) Purification and characterization of an elastinolytic proteinase secreted by cercariae of *Schistosoma mansoni*. *Journal of Biological Chemistry*, 260:3703-3707.

McKerrow, J.H., Brindley, P., Brown, M., Gam, A.A., Staunton, C. and Neva, F.A. (1990) *Strongyloides stercoralis*: identification of a protease that facilitates penetration of skin by the infective larvae. *Experimental Parasitology*, 70:134-143.

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.

Moerman, D.G. and Waterston, R.H. (1984) Spontaneous unstable *unc*-22 *IV* mutations in *C. elegans* var. Bergerac. *Genetics*, 108:859-877.

Moerman, D.G., Benian, G.M. and Waterston, R.H. (1986) Molecular cloning of the muscle gene *unc*-22 in

Caenorhabditis elegans by Tc1 transposon tagging. *Proceedings of the National Academy of Sciences of USA*, 83:2579-2583.

Mori, I., Moerman, D.G. and Waterston, R.W. (1988) Analysis of a mutator activity necessary for germline transposition and excision of Tc1 transposable elements in *Caenorhabditis elegans*. *Genetics*, 120:397-407.

Mount, S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Research*, 10:459-472.

Muller, D., Quantin, B., Gesnel, M.-C., Millon-Collard, R., Abecassis, J. and Breathnach, R. (1988) The collagenase gene family in humans consists of at least four members. *Biochemical Journal*, 253:187-192.

Nagase, H., Enghild, J.J., Suzuki, K. and Salvesen, G. (1990) Stepwise activation mechanisms of the precursor of matrix metalloproteinase 3 (Stromelysin) by proteinases and (4-aminophenyl)mercuric acetate. *Biochemistry*, 29:5783-5789.

Ohkubo, H., Vogeli, G., Mudryi, M., Avvedimento, V.E., Sullivan, M., Pastan, I. and de Crombrugghe, B. (1980) Isolation and characterization of overlapping genomic clones covering the chicken α2 (type I) collagen gene. *Proceedings of the National Academy of Sciences of USA*, 77:7059-7063.

Ohtsuka, E., Matsuki, S., Ikehara, M., Takahashi, Y. and Matsubara, K. (1985) An alternative approach to

deoxyoligonucleotides as hybridization probes by insertion of deoxyinosine at ambiguous codon positions. *Journal of Biological Chemistry*, 260:2605-2608.

Okkema, P. and Kimble, J. (1991) Molecular analysis of *tra*-2, a sex determining gene in *C. elegans*. *EMBO Journal*, 10:171-176.

Orkin, S. (1990) Globin gene regulation and switching: circa 1990. *Cell*, 63:665-672.

Ouazana, R. and Herbage, D. (1981) Biochemical characterization of the cuticle collagen of the nematode *Caenorhabditis elegans*. *Biochimica et Biophysica Acta*, 669:236-243.

Page, A.P. and Maizels, R.M. (1992) Biosynthesis and glycosylation of serine/threonine-rich secreted proteins from *Toxocara canis* larvae. *Parasitology*, 105:297-308.

Page, A.P., Rudin, W., Fluri, E., Blaxter, M.L. and Maizels, R.M. (1992) *Toxocara canis*: a labile antigenic surface coat overlying the epicuticle of infective larvae. *Experimental Parasitology*, 75:72-86.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of USA*, 85:2444-2448.

Politz, S.M., Politz, J.C. and Edgar, R.S. (1986) Small collagenous proteins present during the moult in *Caenorhabditis elegans*. *Journal of Nematology*, 18:303-

310.

Politz, S.M., Philipp, M.T., Estevez, M., O'Brien, P.J. and Chin, K.J. (1990) Genes that can be mutated to unmask hidden antigenic determinants in the cuticle of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of USA*, 87:2901-2905.

Politz, S.M. and Philipp, M.T. (1992) *Caenorhabditis elegans* as a model for parasitic nematodes: a focus on the cuticle. *Parasitology Today*, 8:6-12.

Priess, J. and Hirsh, D. (1986) *Caenorhabditis elegans* morphogenesis: the role of the cytoskeleton in the elongation of the embryo. *Developmental Biology*, 117:156-173.

Probst, J.C., Gertzen, E.-M. and Hoffmann, W. (1990) An integumentary mucin (FIM-B.1) from *Xenopus laevis* homologous with Von Willebrand factor. *Biochemistry*, 29:6240-6244.

Proudfoot, N.J. and Brownlee, G.G. (1976) 3' non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263:211-214.

Robertson, A. and Thomson, J. (1982) Morphology of programmed cell death in the ventral nerve cord of *Caenorhabditis elegans* larvae. *Journal of Embryology and Experimental Morphology*, 67:89-100.

Rose, M.C. (1992) Mucins: structure, function, and role in

pulmonary diseases. *American Journal of Physiology*, 263:L413-L429.

Rosenzweig, B., Liao, L.W. and Hirsh, D. (1983) Sequence of the *C. elegans* transposable element Tc1. *Nucleic Acids Research*, 11:4201-4209.

Rosenquist, T.A. and Kimble, J. (1988) Molecular cloning and transcript analysis of *fem*-3, a sex-determining gene in *Caenorhabditis elegans*. *Genes and Development*, 2:606-616.

Rychlik, W. and Rhoads, R.E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and "in vitro" amplification of DNA. *Nucleic Acids Research*, 17:8543-8551.

Salowe, S.P., Marcy, A.I., Cuca, G.C., Smith, C.K., Kopka, I.E., Hangmann, W.K. and Hermes, J.D. (1992) Characterization of Zinc-binding sites in human Stromelysin-1: stoichiometry of the catalytic domain and identification of a cysteine ligand in the proenzyme. *Biochemistry*, 31:4535-4540.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular cloning, a laboratory manual*. 2nd edition, Cold Spring Harbour (NY): Cold Spring Harbour Laboratory Press.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences of USA*, 74:5463-5467.

Saus, J., Quinones, S., Otani, Y., Nagase, H., Harris, E.D. and

Kurkinen, M. (1988) The complete primary structure of human matrix metalloproteinase-3: identity with stromelysin. *Journal of Biological Chemistry*, 263:6742-6745.

Schaffer, H.E. and Sederoff, R.R. (1981) Improved estimation of DNA fragment lengths from agarose gels. *Analytical Biochemistry*, 115:113-122.

Schenkman, S. and Eichinger, D. (1993) *Trypanosoma cruzi* trans-sialidase and cell invasion. *Parasitology Today*, 9:218-222.

Schenkman, S., Ferguson, M.A., Heise, N., Cardoso de Almeida, M.L., Mortara, R.A. and Yoshida, N. (1993) Mucin-like glycoproteins linked to the membrane by glycosylphosphatidylinositol anchor are the major acceptors of sialic acid in a reaction catalyzed by trans-sialidase in metacyclic forms of *Trypanosoma cruzi*. *Molecular and Biochemical Parasitology*, 59:293-304.

Sebastiano, M., Lassandro, F. and Bazzicalupo, P. (1991) *cut*-1 a *Caenorhabditis elegans* gene coding for a dauer-specific noncollagenous component of the cuticle. *Developmental Biology*, 146:519-530.

Shamansky, L.M., Pratt, D., Boisvenue, R.J. and Cox, G.N. (1989) Cuticle collagen genes of *Haemonchus contortus* and *Caenorhabditis elegans* are highly conserved. *Molecular and Biochemical Parasitology*, 37:73-86.

Shankar, V., Tan, S., Gilmore, M.S. and Sachdev, G.P. (1992)

Molecular cloning of the carboxy terminus of a canine tracheobronchial mucin. *Biochemical and Biophysical Research Communications*, 189:958-964.

Sharp, P.A., Sugden, B. and Sambrook, J. (1973) Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose-ethidium bromide electrophoresis. *Biochemistry*, 12:3055-3063.

Short, J.M., Fernandez, J.M., Sorge, J.A. and Huse, W. (1988) Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Research*, 16:7583-7600.

Singh, R. and Sulston, J. (1978) Some observations on molting in *Caenorhabditis elegans*. *Nematologica*, 24:63-71.

Sorge, J.A. (1988) Bacteriophage lambda cloning vectors. In *Vectors: a survey of molecular cloning vectors and their uses*. R.L. Rodriguez and D.T. Denhardt (eds.), Stoneham (Massachusetts): Butterworth Publishers.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98:503-517.

Spicer, A.P., Parry, G., Patton, S. and Gendler, S.J. (1991) Molecular cloning and analysis of the mouse homologue of the tumor-associated mucin, MUC1, reveals conservation of potential *O*-glycosylation sites, transmembrane, and cytoplasmic domains and a loss of minisatellite-like polymorphism. *Journal of Biological Chemistry*, 266:15099-

15109.

Spieth, J., Denison, K., Kirtland, S. and Cane, J. (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.

Springman, E.B., Angleton, E.L., Birkedal-Hansen, H. and van Wart, H.E. (1990) Multiple modes of activation of latent human fibroblast collagenase - evidence for the role of a cys-73 active-site Zinc complex in latency and a cysteine switch mechanism for activation. *Proceedings of the National Academy of Sciences of USA*, 87:364-368.

Staden, R. (1982) An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Research*, 10:2951-2961.

Staden, R. (1984a) Computer methods to locate signals in nucleic acids. *Nucleic Acids Research*, 12:505-519.

Staden, R. (1984b) Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research*, 12:551-567.

Staden, R. (1984c) Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Research*, 12:521-538.

Staden, R. (1987) Computer handling of DNA sequencing projects. In *Nucleic Acid and Protein Analysis, a Practical Approach*. M.J. Bishop and C.J. Rawlings (eds.), Oxford:

IRL Press.

Starich, T.A., Herman, R.K. and Shaw, J.E. (1993) Molecular and genetic analysis of *unc-7*, a *Caenorhabditis elegans* gene required for coordinated locomotion. *Genetics*, 133:527-541.

Strous, G.J. and Dekker, J. (1992) Mucin-type glycoproteins. *Critical Reviews in Biochemistry and Molecular Biology*, 27:57-92.

Sulston, J. and Horvitz, H.R. (1977) Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. *Developmental Biology*, 56:110-156.

Sulston, J. and Hodgkin, J. (1988) Methods. In *The nematode Caenorhabditis elegans*. W.B. Wood (ed.), New York: Cold Spring Harbor Laboratory.

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., Metztein, M., Hawkins, T., Ainscough, R. and Waterston, R. (1992) The *C. elegans* genome sequencing project: a beginning. *Nature*, 356:37-41.

Tabor, S. and Richardson, C.C. (1987) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proceedings of the National Academy of Sciences of USA*, 84:4767-4771.

Takahashi, Y., Kato, K., Hayashizaki, Y., Wakabayashi, T.,

Ohtsuka, E., Matsuki, S., Ikehara, M. and Matsubara, K. (1985) Molecular cloning of the human cholecystokinin gene by use of a synthetic probe containing deoxyinosine. *Proceedings of the National Academy of Sciences of USA*, 82:1931-1935.

Tautz, D. and Renz, M. (1983) An optimized freeze-squeeze method for recovery of DNA fragments from agarose gels. *Analytical Biochemistry*, 132:14-19.

Thomas, M. and Davis, R.W. (1975) Studies on the cleavage of bacteriophage lambda DNA with EcoRI restriction endonuclease. *Journal of Molecular Biology*, 91:315-328.

van Luenen, H.G.A.M., Colloms, S.D. and Plasterk, R.H.A. (1993) Mobilization of quiet, endogenous Tc3 transposons of *Caenorhabditis elegans* by forced expression of Tc3 transposase. *EMBO Journal*, 12:2513-2520.

van Wart, H.E. and Birkedal-Hansen, H. (1990) The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences of USA*, 87:5578-5582.

von Mende, N., Bird, D.M., Albert, P.S. and Riddle, D.L. (1988) *dpy*-13: a nematode collagen gene that affects body shape. *Cell*, 55:567-576.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metztein, M., Hawkins, T., Wilson, R.,

Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J. and Sulston, J. (1992) A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genetics*, 1:114-123.

Way, J.C. and Chalfie, M. (1988) *mec*-3, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*. *Cell*, 54:5-16.

White, J. (1988) The anatomy. In *The nematode Caenorhabditis elegans*. W.B. Wood (ed.), Cold Spring Harbour (NY): Cold Spring Harbour Laboratory Press.

Whitham, S.E., Murphy, G., Angel, P., Rahmsdorf, H.J., Smith, B., Lyons, A., Harris, T.J.R., Reynolds, J.J., Erlich, P. and Docherty, A.J.P. (1986) Comparison of human stromelysin and collagenase by cloning and sequence analysis. *Biochemical Journal*, 240:913-916.

Wood, W.B. (1988) The nematode Caenorhabditis elegans. Cold Spring Harbour (NY): Cold Spring Harbour Laboratory Press.

Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and Doty, P. (1981) Structure of the pro-α2 (I) collagen gene. *Nature*, 294:129-135.

Wright, K.A. (1987) The nematode's cuticle - its surface and the epidermis: function, homology, analogy - a current consensus. *Journal of Parasitology*, 73:1077-1083.

Xu, G., Huan, L.J., Khatri, I., Wang, D., Bennick, A., Fahim, R.E., Forstner, G.G. and Forstner, J.F. (1991) cDNA

for the carboxyl terminal region of the rat intestinal mucin-like peptide. *Journal of Biological Chemistry*, 267:5401-5407.

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.

Zuckerman, B.M., Kahane, I. and Himmelhoch, S. (1979) *Caenorhabditis briggsae* and *C. elegans*: partial characterization of cuticle surface carbohydrates. *Experimental Parasitology*, 47:419-424.