Interpretation of the *Caenorhabditis elegans* genome sequence data through gene expression patterns.

Archana Sharma-Oates

Submitted in accordance with the requirements for the degree of Doctor of Philosophy.

The University of Leeds
School of Biology.

July 2003.

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgments.

I would like to thank my supervisor, Dr I. A. Hope for his time and patience and for providing me with a valuable insight into academic scientific research. Many thanks to Dr D. R. Westhead for help and advice on the bioinformatic aspect of this research, Dr S. Pickering for his technical assistance and Dr A. Mounsey for many helpful discussions. In addition, particular thanks to Professor Y. Kohara at the National Institute of Genetics, Mishima, Japan for his invitation to spend two months analyzing gene expression patterns in his laboratory. I am also grateful to my friends Nicola Gold, Petra Bauer, Andrew Nightingale, Vidhya Gomathi Krishnan, Sid (Michael Sadowski) and all members of the bioinformatics group for their help and advice over the last three years.

On a personal note, special thanks to my husband, Adam for his continual support and encouragement, as well as his proof-reading of this thesis and to my daughter Anya for putting up with me during the writing of this thesis. In addition I would like to thank family and, in particular, my parents-in-law for numerous discussions on the function of genes.

Abstract.

The nematode *Caenorhabditis elegans* has been studied extensively as a means of understanding development and cellular processes and was the first multicellular organism to have a sequenced genome. A number of *C. elegans* gene expression patterns have been characterized, using several different experimental approaches, thereby providing a link between the nucleic acid sequence of a gene and the temporal and spatial nature of its expression. A systematic collation and analysis of *C. elegans* gene expression pattern data revealed a high degree of agreement in the results obtained using the different experimental approaches. During this analysis a group of genes was identified that expressed specifically in one particular cell type, the excretory cell. To develop a strategy to identify *cis*-acting regulatory elements responsible for the control of cell type-specific expression, the DNA sequences of the potentially co-regulated excretory cell-expressing genes were analysed using two software packages, MEME and SPEXS.

The MEME output contained many DNA motifs but their sequence simplicity suggested that they were unlikely to be genuine regulatory elements. In contrast, the output from SPEXS identified a vast number of more complex motifs. However because SPEXS detects motifs simply based on sequence, without considering biological characteristics of *cis*-acting elements, no priority was assigned to the identified elements. Therefore a scoring strategy was devised that incorporated different weightings such that motifs occurring with high frequency within 1 kb upstream from the translational start, and with high sequence complexity, were assigned a higher score.

To test the effectiveness of the scoring strategy when applied to the SPEXS output, a *C. elegans* muscle data set was analysed which was known to contain a previously characterized *cis*-acting element. The element in question was identified suggesting that the scoring strategy worked well. When the strategy was applied to excretory cell data set the highest scoring motif, and therefore most likely candidate *cis*-acting element, was the motif TTACCGAA. This motif was also detected in test sets containing excretory cell-expressing genes and a data set containing *C. briggsae* orthologues of *C. elegans* genes. These results suggest that the scoring strategy is an effective approach to identify *cis*-acting elements and the motif TTACCGAA is potentially such an element which mediates excretory cell expression in *C. elegans*.

Contents.

List of Tables.

List of Figures.

Abbreviations.

ACeDB	A *C.elegans* database
BLAST	Basic local alignment search tool
bp	Base pair
C .briggsae	*Caenorhabditis briggsae*
cDNA	Complementary deoxyribonucleic acid
C. elegans	*Caenorhabditis elegans*
DNA	Deoxyribonucleic acid
Drosophila	*Drosophila melanogaster*
EBI	European bioinformatics institute
E. coli	*Escherichia coli*
EM	Expectation maximization
EST	Expressed sequence tags
FASTA	Fast alignment search tool algorithm
GFP	Green fluorescent protein
kb	Kilobases pairs
Mb	Megabases
MEME	Multiple expectation-maximization for motif elicitation
-mod	Model of distribution of motifs
mRNA	Messenger ribonucleic acid
NCBI	National centre for biotechnology information
-neg	negative set option
-nmotifs	Number of motifs
nt	Nucleotide
OODBMS	Object-orientated database management system
OOPS	One occurrence per sequence
OST	Open-reading-frame sequence tag
RNA	ribonucleic acid
RT-PCR	Reverse transcriptase polymerase chain reaction
Saccharomyces	*Saccharomyces cerevisae*
SGD	*Saccharomyces* genome database
SPEXS	Sequence pattern exhaustive search

TCM	Two component mixture
TFs	Transcription factors
URLs	Uniform resource locators
-W	Motif width
WWW	World wide web
X-gal	5-bromo-4-chloro-3-indoyl-β-D galactoside
YAC	Yeast artificial chromosome
ZOOPS	Zero or one occurrence per sequence

Chapter 1

Chapter 1: General introduction.

1.1: *Caenorhabditis elegans* as a model organism.

Since the pioneering work of Sydney Brenner in the 1960's, the free-living soil nematode, *Caenorhabditis elegans* (*C. elegans*) has been extensively studied and has proved an excellent model to study numerous aspects of developmental and cellular biology. In addition it was the first multicellular organism to have the complete nucleic acid sequence determined and has led the field of post-genomics research (Chalfie, 1998).

There are a number of advantages inherent to *C. elegans* that make it a particularly suitable model system that may shed light on cellular processes in more complex organisms e.g. humans. *C. elegans* is a simple organism both anatomically and genetically. It exists in two distinct sexes the hermaphrodite and the male. The adult hermaphrodite and adult male have 959 and 1031 somatic nuclei, respectively and the haploid genome size is 100 Mb (Plasterk, 1996). *C. elegans* is easily maintained in the laboratory and large numbers can be grown in mass culture on agar plates or in liquid culture using *Escherichia coli* as a food source (Hodgkin *et al.*, 1995). Individual animals can be easily observed and manipulated with the aid of a dissecting microscope. As the animals are transparent throughout their life cycle, organism development can be studied at the "living", cellular level by light microscopy. In addition, its small size allows complete anatomical description of the animal at the electron microscope level (Hodgkin *et al.*, 1995). Mutants are readily obtained following chemical mutagenesis or exposure to ionizing radiation. The simplicity, convenience of manipulation, and short life cycle of *C. elegans* in addition to the high degree of conservation with human biological processes make it an excellent experimental organism for the study of development (Blaxter, 1998).

C. elegans naturally thrives in many parts of the world and in optimal conditions reproduces with a life cycle of approximately 3 days (Plasterk, 1999). The two sexes, hermaphrodites and males, are *circa* 1 mm in length but differ in appearance as adults (Figure 1.1). Hermaphrodites produce both oocytes and sperm, and can reproduce by self-fertilization. Males, which arise spontaneously at low frequency, can fertilize

hermaphrodites (male sperm have a competitive advantage over hermaphrodite sperm); hermaphrodites cannot fertilize each other (Chisholm and Jin, 2001).

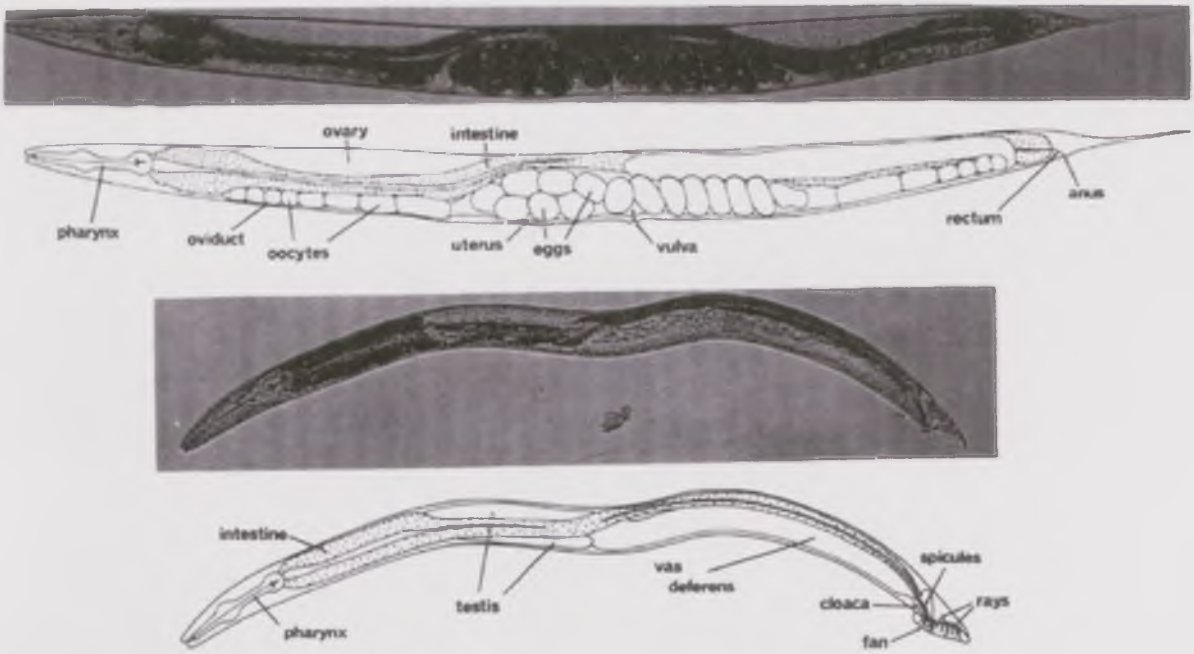

Figure 1.1. Photomicrographs showing major anatomical features of the *C. elegans* adult. Hermaphrodite (top) and male (bottom). Shown are lateral views under bright-field illumination and labelled line drawings to indicate anatomy. (Figure taken from Suslton and Horvitz, 1977; Riddle *et al.*, 1997).

An unmated hermaphrodite releases in the region of 300 eggs during its reproductive life. Juvenile worms hatch and develop via four distinct stages (commonly referred to as larval stages, although no metamorphosis is involved), separated by molts (Plasterk, 1999). The mature adult emerging from the fourth molt is fertile for approximately 4 days and then lives for a further 10-15 days.

1.2: *C. elegans* anatomy.

C. elegans has a body with an outer tube that consists of cuticle, hypodermis, neurons, and muscles surrounding a pseudocoelomic space that contains the intestine and gonad (Kimble and Hirsh, 1979; Seydoux and Strome, 1999; Vogel and Hedgecock, 2001). The shape of the worm is maintained by internal hydrostatic pressure, controlled by an osmoregulatory system (White, 1988; Buechner, 2002). The three-layered collagenous cuticle is secreted by the underlying hypodermis. This tissue is syncytial (made up of

large multinucleate cells). In adults, lateral, longitudinal cords of seam cells form treads (alae) on the cuticle surface (Johnstone, 2000). On solid media the worm crawls on one side, with the alae contacting substrate.

The obliquely striated body-wall muscle cells of *C. elegans* are arranged into four strips running the length of the animal, two dorsally and two ventrally. Most of the cells of the nervous system are found surrounding the pharynx, along the ventral midline, and in the tail. Processes from these neurons form an external ring around the pharynx (the nerve ring) or contribute to process bundles running the length of the body, the most prominent being the dorsal and ventral nerve cords (Hall and Russell, 1991). Sensory neurons run anteriorly from the nerve ring to sensory organs (sensilla) in the head. The nerve ring receives inputs from the head region and sends an output primarily to the body-wall muscles via motor neuron axons in the ring itself and in the dorsal and ventral cords (White, 1988).

C. elegans feeds through a bi-lobed pharynx, which compresses food into the intestine, crushing it as it passes through the second lobe (Albertson and Thomson, 1976). The intestine is formed from two rows of eight cells plus an anterior ring of four, surrounding a central lumen, which connects near the tail to the anus (Sulston and Horvitz, 1977). The simple excretory system is probably also responsible for osmoregulation (Bürglin and Ruvkun, 2001). It consists of a pair of excretory canals, which are processes of a single cell that run the length of the animal, connecting to the exterior through the anteriorly located excretory pore. Due to the simplicity of the excretory cell, this cell type was selected for analysis and is of fundamental importance for the research described later in this thesis and the identification *cis*-acting elements (Section 1.10).

In a project spanning a decade, John White and colleagues described the anatomy of the *C. elegans* nervous system. By assembling thousands of electron micrographs of serial sections, 302 neurons (compared to 100 billion or so in humans) and their associated connections were identified. In 1986 these findings ("known as the wiring diagram") were published as a single, 340-page issue of the *Philosophical Transactions of the Royal Society of London* (White *et al.*, 1986; Chalfie, 1998). In a second large-scale project, John Sulston and his colleagues described the complete cell lineage of *C.*

elegans, commencing with the zygote and tracking each cell division ultimately to the 959 somatic cells in the adult hermaphrodite and 1031 in the male. As a result of these studies, the complete cellular arrangement of *C. elegans* has been described and every cell in the animal has been assigned a unique label (Sulston and Horvitz, 1977).

1.3: Life stages.

1.3.1: Embryogenesis.

Embryogenesis in *C. elegans,* from fertilization to hatching, takes approximately 14 hours at 22°C (Figure 1.2). The process can be conveniently considered in three major stages (Deppe *et al.*, 1978; Sulston *et al.*, 1983). The first stage, which includes zygote formation and early cleavage, establishment of the embryonic axes, and determination of the somatic and germline founder-cell fates, takes place in the first two hours after fertilization. The second stage, of gastrulation, completion of most cell proliferation, and the beginning of cell differentiation and organogenesis, continues until approximately midway through embryogenesis. The third stage of morphogenesis, as well as completion of embryonic cell differentiation and organogenesis, consists of the remainder of embryogenesis and concludes with hatching (Sulston *et al.*, 1983).

1.3.2: Postembryonic development.

After 14 hours the embryo hatches to become the first stage (L1) larva. *C. elegans*, like many nematodes, has four larval stages (L1-L4). During postembryonic development the animal increases from 250 μm to 1mm in length (Sulston and Horvitz, 1977; Chisholm and Jin, 2001). Cell proliferation occurs in the epidermis throughout larval development, and in the peripheral nervous system (ventral cord) in the L1 and L2 stages. Sexual maturation occurs during the L3 and L4 stages, and involves the growth of the somatic gonad, proliferation of the germline (Kimble and Hirsh, 1979), and formation of the hermaphrodite vulva and the male copulatory apparatus (tail).

The four larval stages are separated by molts, in which the previous cuticle is shed and a new cuticle secreted. The L4 larva molts to become an adult (Figure 1.2) (Chisholm and Jin, 2001).

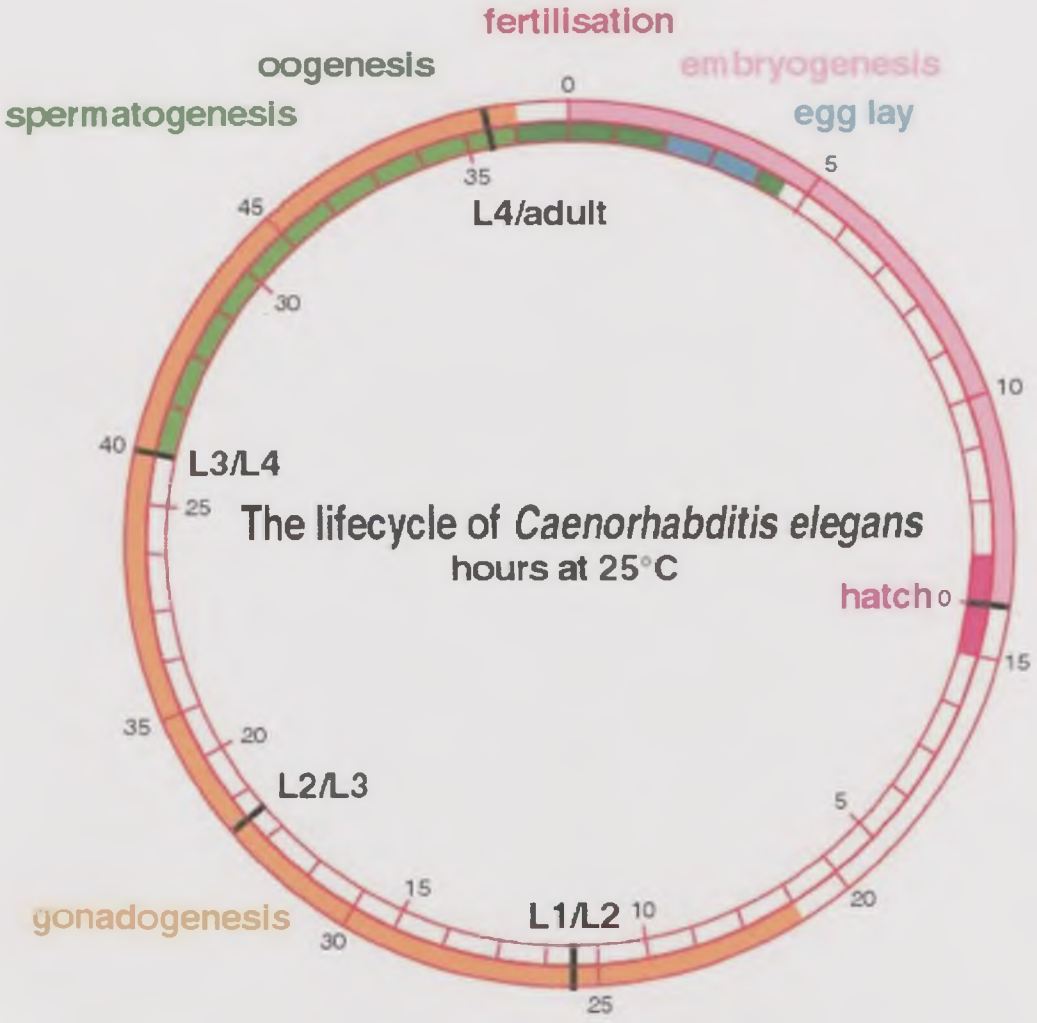

Figure 1.2. Schematic representation of the *C. elegans* life cycle. The outer and inner rings represent time (hours) after fertilization and hatching, respectively. (Figure from Wood *et al.*, 1980 and Mark Blaxter's Homepage [http://nema.cap.ed.ac.uk/Caenorhabditis/C_elegans.html]).

1.4: The *C. elegans* genome project.

1.4.1: The physical map.

To further develop the understanding of *C. elegans*, in the 1980s Coulson, Sulston, Waterston and colleagues initiated the process of developing a clone-based physical map of the *C. elegans* genome (Coulson *et al.*, 1986). The map was initially based on cosmid clones mapped using a fingerprinting approach. Initially restriction enzymes were used to digest the chromosomes into fragments which were then cloned into cosmid vectors (Coulson *et al.*, 1991). The fragments were then reassembled using a fingerprinting strategy that involved the computer-aided identification of distinctive patterns or fingerprints within individual clones. Clones with overlapping sequences

were then assembled such that a long sequence of contiguous clones was generated. This map consisted of 700 separate contigs with, 700 intervening gaps (Roberts, 1990; Wilson, 1999). To fill-in these gaps, larger fragments were then cloned into yeast artificial chromosome (YAC) vectors and again overlapping clones were detected (Coulson *et al.*, 1988). As a consequence of these efforts the number of gaps was reduced to 150 and the YACs provided coverage of approximately 20% of the genome not represented in the original cosmid libraries (Roberts, 1990). Through a combined international effort, the entire genome has been cloned and all the sequences have been assigned to individual chromosomes (Waterston and Sulston, 1995).

1.4.2: The sequencing project.

With the completion of the physical map, the feasibility of sequencing the entire 100 Mb (http://www.sanger.ac.uk/) of the genome became apparent (Sulston *et al.*, 1992). Initially the approach for sequencing was primer-directed or 'walking' using the cosmid clones as template for sequencing reactions (Berks *et al.*, 1995; Wilson, 1999). However, this approach was quickly abandoned, largely due to its relative inefficiency. In an effort to increase sequencing throughput, a robotic system was introduced (Watson *et al.*, 1993). In addition a number of biochemical procedures were developed to allow the automation of a shotgun sequencing protocol. The first step in this new strategy was a plaque-picking procedure to select individual M13 clones, followed by the growth in culture. The next step was purification of the DNA template and then sequencing of each of M13 clone before being loaded on to the DNA sequencer (Watson *et al.*, 1993).

Pursuing this high throughput approach the genome sequencing project was largely completed by 1998 with accuracy of the final sequenced product thought to be less than one error per 10,000 bases (The *C. elegans* sequencing consortium, 1998).

1.4.3: The ACeDB database.

To provide computational support to the *C. elegans* sequencing project, the database, ACeDB was developed by Jean Thierry-Meig and Richard Durbin. This object-orientated database management system (OODBMS) has built-in routines for handling DNA and protein sequences, genome maps, and other biological entities. It can be used as a "standalone" application which comes complete with graphical displays for many

specialised biological data, or as a multiuser client-server system. In addition to the graphical interface, ACeDB has a text-only query-language interface, as well as programmers' interfaces written in C, Perl, and Java. ACeDB is also supported for Unix, Windows and Mac operating systems, and it is an open source project, with the entire source code available for examination and modification (Stein and Thierry-Meig, 1999). ACeDB contains the complete cell lineage of *C. elegans*, genetic maps, strain and phenotype information from mutant studies, information on gene expression patterns and extensive bibliography that includes unpublished abstracts and short communications (Stein, 1999).

The ACeDB database is available at anonymous ftp servers from the Sanger Centre, Washington University and the NCBI. The ACeDB database has been applied to several other genome projects such as *S. cerevisiae* (SGD, Cherry *et al.*, 1998), the human genome at the Sanger Centre and the Washington University Genome Sequencing Centre (Waterston and Sulston, 1995), Xenopus and the flowering plant *Arabidopsis thaliana* (Chalfie, 1998; Stein and Thierry-Meig, 1998).

The web-based version of ACeDB, known as WormBase (Stein *et al.*, 2001) is a collaborative effort by laboratory groups in USA and the Sanger Centre, UK. It is continually being improved (Harris *et al.*, 2003) and updated to capture, curate and distribute data on *C. elegans* biology. It builds upon the existing ACeDB database by providing data curation services and aims to make the database more "user-friendly". In addition, the WormBase database is available for bulk download and is not subject to any license restrictions.

1.4.4: Findings from the sequencing project.

When completed in 1998, the *C. elegans* sequencing project was the first multicellular organism to have its entire genome sequenced. The complete *C. elegans* genome sequence consists of 100264085 bp of DNA. Although there is some conjecture the 100 Mb of DNA encodes between 19,000 to 19,757 predicted genes (Ashrafi *et al.*, 2003; Kamath *et al.*, 2003; Tuschl, 2003). Precisely how many of these genes are genuine has been assessed by Reboul *et al.* (2001) using an OST (open-reading-frame sequence tag) approach. They selected 1,222 predicted genes for which no EST (expressed sequence tag) had previously been obtained, and attempted to amplify a predicted product from

cDNA. Using this approach in excess of 70% of these genes were identified as *bona fide*, although the predicted intron/exon structure was not always correct (Reboul *et al.*, 2001; Hodgkin, 2001). Approximately 50% of the *C. elegans* genes are currently enigmatic in terms of sequence similarity to other species and function, however, it has been suggested that many of these sequences could be pseudogenes (Mounsey *et al.*, 2002).

The coding sequence accounts for 27 % of the genome, with introns (an average of five per gene) accounting for a further 26 %. The five autosomal chromosomes have a greater gene density in the center than on chromosomal arms. The X chromosome does not show regional differences. Matches to *C. elegans* complementary DNAs and to non-nematode proteins are also greatest in the central region, whereas tandem and inverted repeats are more common on the arms. These findings suggest that the chromosomal arms may be areas of rapid evolutionary change (Chalfie, 1998).

A comparison between human and *C. elegans* predicted proteins indicates that 32% of *C. elegans* proteins are similar to human sequences, whereas 70% of human proteins identify similar sequences in *C. elegans*. The worm, however, lacks genes for some proteins, such as sodium channels, trk receptors and connexins (Chalfie, 1998).

1.5: Gene expression patterns.

Since the completion of the *C. elegans* genome sequencing project a major challenge has been to define the role of each of the predicted genes. Large-scale studies have been pursued to understand gene function using a gene silencing strategy by high-throughput RNAi experiments (Fraser *et al.*, 2000; Maeda *et al.*, 2001; Kamath *et al.*, 2003). An additional experimental approach that can be used to study how DNA sequence information directs the 4-dimensional development of the animal is through the analysis of gene expression patterns (Hope *et al.*, 1996). Understanding the temporal and spatial nature of the expression of a gene often provides important evidence as to its biological role. Genes with similar expression patterns are potentially co-regulated and therefore provide circumstantial evidence linking genes and pathways to particular phenotypes and biological processes (DeRisi *et al.*, 1997; Eisen *et al.*, 1998; Heyer *et al.*, 1999; Holter *et al.*, 2000; Mir, 2000). The approaches used to characterize gene expression patterns are reporter gene fusion, *in situ* hybridization,

immunostaining and expression profiles. The latter of which are primarily monitored using microarray technology as well as western, northern and RT-PCR analyses, (all techniques are described in detail in Chapter 3). The *C. elegans* organism is particularly well-suited to gene expression studies because throughout development most somatic cells can be identified *via* light microscopy and the precise location of gene expression can be determined (Sulston and Horvitz, 1977; Kimble and Hirsh, 1979; Sulston *et al.*, 1980; Sulston *et al.*, 1983).

1.6: Transcription factors.

Transcription factors (TFs) have long been recognized to play a major role in regulating gene expression in eukaryotic organisms. Specific TFs are produced in the cell in response to intercellular signaling and also asymmetric cell division. By binding to sequence-specific sites (Grabe, 2000; Vilo *et al.*, 2000) (known as transcription factor binding sites or *cis*-acting elements) located in promoter regions of genes, TFs influence the transcription of a particular gene. A substantial aspect of this regulation can be attributed to the interaction of TFs with specific *cis*-acting regulatory DNA sequences. These regulatory sequences are arranged as distinct units (enhancers), with each unit containing one or more *cis*-acting elements for a specific combination of TFs. The combinatorial binding of TFs facilitates the expression of genes in a particular developmental context and provides the tight spatial and temporal regulation of gene transcription that is vital for the development of the organism.

Regulatory elements often exhibit considerable variability in DNA sequence thereby providing the opportunity for subtle transcriptional control. This is of obvious importance as some proteins are required at much higher levels than others. Similarly, regulatory proteins often control the expression of a number of genes which are expressed at different levels. This too can be achieved by elements having variations in sequence and therefore different affinities for regulatory proteins.

The identification of *cis*-acting regulatory elements is an important step in understanding the mechanism of transcriptional regulation of a particular gene. Genes with similar expression patterns may be transcriptionally co-regulated, and their regulatory regions can be analysed for the presence of common DNA sequence motifs (Pickert *et al.*, 1998; Bucher, 1999). However, the *de novo* detection of *cis*-acting

regulatory elements is a difficult task as the element is of unknown size, may be poorly conserved between promoters and the sequence used to search for the motif may not represent the complete promoter region (Klingenhoff *et al.*, 1999; Kolchanov *et al.*, 1999; Ohler and Niemann, 2001). In addition several regulatory mechanisms may lead to the same expression pattern and co-expression of a group of genes does not necessarily mean co-regulation (Vilo *et al.*, 2000).

1.7: Computer algorithms for analysis and prediction of *cis*-acting regulatory elements.

1.7.1: The identification of *cis*-acting elements.

The recent availability of the entire genome DNA sequence has created the potential for identifying *cis*-acting regulatory elements via bioinformatic approaches (Halfon *et al.*, 2002). Due to the difficulty in identifying *cis*-acting elements, studies have mainly concentrated on the relatively 'simple' genome of the budding yeast *Saccharomyces cerevisiae* (Brazma *et al.*, 1998a; van Helden *et al.*, 1998). In this organism most of the known regulatory elements are close to the translational start of the genes, the majority being found 10-700 base pairs (bp) upstream from the translation start codon. Therefore in the case of yeast, the 1 kilobase (kb) region upstream of the start codon can be used as a good predictor of the location of a promoter region. Therefore the majority of the algorithms searching for conserved motifs in yeast promoters use as the data set the region of 500-1000 bp upstream of the translational start of potentially co-regulated genes. A number of different algorithms have been developed for the detection of *cis*-acting elements. These can be divided into alignment-based approaches and enumerative-based approaches. The first alignment based approach using a multiple alignment strategy was the CONSENSUS algorithm. This algorithm generates a multiple alignment of sequences by aligning them one by one until the information content of the weight matrix constructed from the alignment is at an optimum (detailed discussion in Sections 1.7.2 and 1.7.3). Other alignment based algorithms use a statistical approach; they consider the start positions of the motifs in the sequences to be unknown and perform a local optimization to determine which positions deliver the most conserved motif. Two important methods of this type are Gibbs sampling and expectation maximization in the MEME (Multiple Expectation-maximization for Motif Elicitation) system (discussed in detail in Chapter 4).

An alternative approach to the alignment strategies discussed above is the enumerative, also known as exhaustive, method which the SPEXS (Sequence Pattern EXhaustive Search) algorithm is an example of. These algorithms examine all motifs up to a certain length and report those that occur more frequently than expected based on the overall promoter sequence composition (for further details see Chapter 5). From a practical standpoint, the principle difference between the weight matrix (alignment) method and the enumerative method is the presentation of the result: the CONSENSUS and the MEME system generate a model of the motifs (usually a weight matrix) built from the alignment, whereas the enumerative methods give a list of all motifs (Ohler and Niemann, 2000).

Strategies that have been successfully used in yeast are difficult to extend to higher eukaryotes where the regulatory modules are extensive and can be located many kb either side of a coding region or within an intron (Gaudet and Mango, 2002). Other methods rely on models derived from the prior characterization of a large number (ten or more) of regulatory elements of similar function, however for most genes such extensive information is not available. The identification of dense clusters of known *cis*-acting elements has also been used as the basis for computational searching. However, the predictive value of these various approaches remains uncertain and although they have been successful at recognizing known *cis*-acting elements, little experimental validation of the identification of putative novel elements has been performed (Halfon *et al.*, 2002).

1.7.2: Identifying *cis*-acting elements using a consensus approach.

The derivation of a consensus sequence has been widely used as a means to represent *cis*-acting elements. A consensus sequence refers to a sequence that closely matches the sample motifs, but not necessarily exactly (Figure 1.3). Defining a consensus sequence can be problematic and there is compromise to be reached between the number of mismatches allowed and the sensitivity and precision of the consensus sequence. Although a derived consensus sequence may represent a collection of motifs, it is difficult to define a sequence that can be used to search for the occurrence of new motifs within a sequence (Stormo, 2000). With the aim of identifying *cis*-acting elements in co-regulated genes (from expression studies) there is a collection of sequences that are known to contain *cis*-acting elements for a common factor, but

neither the positions of the *cis*-acting elements nor the specificity of the factor are known.

The use of consensus sequences for motif identification began during studies to sequence *E. coli* promoter regions. From those few sequences the –10 and the –35 consensus sequences were determined 'by direct examination'. This approach was possible because there were only a few sequences and they could be aligned (approximately) because the start of transcription was known. It was observed that all of the sequences had very similar motifs at two locations, approximately 10 and 35 bases upstream of the start. However, as more sequences were collected and with little information available as to the alignment, computer algorithms were required to locate the important features (Queen *et al.*, 1982; Stormo 2000). The first such algorithm was developed by Galas *et al.* (1985) in which a search was performed for common 'words' (of user specified length) and their 'neighbours', (i.e approximate matches to those words) over a window of possible alignments.

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
TATAAT consensus sequence
TATRNT alternative consensus sequence

Figure 1.3. Alignment of the –10 region of the six promoters from Pribnow (1975). Two possible consensus sequence representations obtained from the alignment.

Within recent years a number of algorithms have been designed to identify consensus sequences from unaligned DNA sequences (Frech *et al.*, 1993; Ulyanov and Stormo, 1995; Quandt *et al.*, 1995; Wolfertsteter *et al.*, 1996; Scherf *et al.*, 2000). These consensus methods have been applied to collections of yeast genes that, based on expression array analysis, are known (or expected) to be co-regulated (van Helden *et*

al., 1998; Brazma *et al.*, 1998a). These methods have been successful at identifying the correct *cis*-acting elements on control data sets, where the *cis*-acting elements are known to occur, however there are not a sufficient number of known *cis*-acting elements to rigorously assess the efficacy of these algorithms. In addition, very few of the predicted motifs have been confirmed experimentally. A further disadvantage of these algorithms is that they are limited to the detection of relatively simple motifs that are short in length with a highly conserved core (van Helden *et al.*, 1998).

1.7.3: Searching for *cis*-acting elements with a weight matrix.

An alternative to deriving a consensus sequence is to directly search for a weight matrix that can differentiate between the sequences known to be co-regulated and other mostly, unregulated sequences (Figure 1.4). A weight matrix is constructed from the alignment of all the motifs and is a log-odds matrix calculated by taking the log (base 2) of the ratio of the probability of a particular letter in the motif occurring at that position, and the average frequency of that letter. There is a numerical value for all possible bases at every position in the motif. The score for any particular motif is the sum of these values for each optimal letter within the motif (Figure 1.4). Any sequence that differs from the consensus motif represented by the weight matrix will have a lower score, but the level of decrease depends on the degree of differences.

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-48
T	17	-32	8	-9	-6	19

Figure 1.4. **Weight matrix representation for the –10 region of *E. coli* promoters.** The elements in red correspond to the consensus sequence TATAAT, with a score of 85. The matrix is based on a large collection of –10 regions (Stormo 2000).

This is a convenient method for representing positions that are more highly conserved than others. There are several issues that need to be addressed with the consensus approach. These include, determining the threshold beyond which the predicted *cis*-acting elements are no longer considered genuine and also the sensitivity and precision with which the *cis*-acting elements are predicted. In addition, the major issue with the weight matrix methods is how to select the elements of the matrix to represent the

motif. Several methods have been proposed to overcome this problem (discussed below) and a number of efficient methods exist for calculating the distribution of scores that can be used to determine statistically significant matches (Horten and Kanehisa, 1992; Claverie and Audic, 1996; Stormo, 2000).

1.7.4: Methods used to determine the weights for the matrix.

One of the first methods used to determine the appropriate values of the weights for the matrix was using a simple neural network machine learning algorithm known as the 'Perceptron' (Stormo *et al.*, 1982). The machine learning algorithm is trained on the examples of known motifs, and motifs that are non specific, to derive a matrix and a threshold that distinguishes the two sets. The weight matrix is then used to search for new motifs from unseen data. The matrix method has been shown to be more sensitive and precise than the best consensus method available (Stormo *et al.*, 1982).

Alternative methods for obtaining weight matrices from purely statistical analyses of *E. coli* promoters have also been developed. A method introduced by Staden (1984) was similar to the above algorithm although it did not allow insertions and deletions within the motifs. In this method the weights are simply the natural logarithms of the frequencies of each base at each position. Therefore the sum of any particular motif is the negative logarithm of the probability of observing that particular sequence in the collection of known motifs (assuming the positions are independent).

1.8: Algorithms incorporating a statistical approach to identify *cis*-acting elements.

A statistical approach for the detection of *cis*-acting element was first developed using a "greedy" algorithm by Hertz *et al.* (1990). The algorithm builds up an entire alignment of the motifs by adding in a new motif at each iteration. The best alignment of potential motifs is considered to be the one with the highest information content (Iseq) calculated using equation 1 shown below.

$$I_{seq}(i) = \sum_{b} fb;i \log_2 \frac{fb;i}{\rho b}$$

Equation 1. Where i is the position within the motif, b refers to each of the possible bases, and fb,i is the observed frequency of each base at the position. ρb is the frequency of base b in the whole genome, with $\rho b = 0.25$ (although this value would be different due to *C. elegans* genome being AT-rich) for all b (Stormo, 1990).

the whole genome, with $\rho b = 0.25$ (although this value would be different due to *C. elegans* genome being AT-rich) for all b (Stormo, 1990).

Recent advances have enabled the calculation of a p-value for each of the alignments which is then used to rank different alignments (Hertz and Stormo, 1999). An expectation maximization (EM) method was developed for the same problem by Lawrence and Rielly (1990). The EM algorithm performs supervised learning and takes as input a set of unaligned sequences and a motif length, and returns a probabilistic model of a motif common to all the sequences. It is essential that each sequence must contain an example of the motif and assumes that the start position of the motif is unknown. The EM approach can be described as an iteration between two steps, for obtaining global maximum likelihood parameter estimates for a model of observed data: a collection of sites of length k whose positions are not known in a set of unaligned sequences, one site per sequence. The resulting collection corresponds to a matrix. The algorithm involves calculating expected parameter values that describe the data (i.e., the matrix of putative sites), then maximizing the likelihood of obtaining these values (i.e., refining the matrix by maximizing its relative information content). The two steps are repeated until the parameters that best describe the data are obtained, or until a fixed maximum number of iterations is reached (Vanet *et al.*, 1999).

As shown in Figure 1.4, given a weight matrix of unaligned sequences it is possible to calculate the score for all possible motifs of specified length in each of the sequences. Using that score, a weighted alignment of all the possible motifs can be obtained. This alignment is used to derive a new matrix representation for those motifs. These two steps are repeated for each possible motif until a motif common to all sequences is found. However, this approach does not always find the optimal alignment of motifs. Bailey, Grundy and Elkan have also developed an EM approach to this problem (Bailey and Elkan, 1994; Bailey and Elkan, 1995a; Grundy *et al.*, 1996), which is implemented in the MEME program. The MEME method allows for the simultaneous identification of multiple motifs, discussed further in Chapter 4. Lawrence and colleagues also developed a 'Gibbs Sampling' variation of the EM method (Lawrence *et al.*, 1993) which, has also been used to define weight matrices for known transcription factors (Schug and Overton, 1997). Zhang recently developed a version of the EM method and used it on sets of co-regulated yeast genes (Zhu and Zhang, 1999) and a similar approach has been used on *E. coli* (Robinson *et al.*, 1998).

In the majority of methods to identify the weight matrix directly from unaligned sequences, the best alignment has the highest information content (overall conservation). This compensates for the base composition of the genome and often identifies the element being sought. However, the assumption of a random genome can be too poor an estimate, and instead of finding the correct elements the methods identify some other motifs that appear to be significant but do not discriminate between the promoters in the collection. For example, many yeast promoters have unexpectedly common stretches of poly(A) or poly(T) sequences, which can appear as motifs identified by the programs. But those motifs occur in many promoters, not just the subset known to be co-regulated, and so cannot be the motif of interest (Workman and Stormo, 2000). This problem has not been successfully overcome by any of the algorithms described in this Section.

1.9: Phylogenetic footprinting.

Recently the use of "phylogenetic footprinting" has gained in popularity for identifying *cis*-acting elements. The simple premise underlying "phylogenetic footprinting" is that selective pressure causes functional elements to evolve at a slower rate than nonfunctional sequences (Wasserman *et al.*, 2000). This implies that well conserved DNA motifs among a set of orthologous promoter regions are excellent candidates for functional *cis*-acting regulatory elements. This approach has proved successful in the identification of regulatory elements for many genes, including *interleukin* (IL)-4, IL-13 and IL-5 that have been experimentally confirmed (Duret and Bucher, 1997; Loots *et al.*, 2000). The major advantage of "phylogenetic footprinting" over a single genome co-regulated multi-gene approach, is that it is capable of identifying regulatory elements specific to a single gene, as long as they are sufficiently conserved across many of the species considered (Blanchette and Tompa, 2002). However, the caveat to this approach is that when including distantly related sequences, there is an increased possibility that some elements may have been altered or lost over the course of evolution. For example, a species may no longer need the regulatory mechanism in which some regulatory element was involved, in which case the selective pressure would no longer apply. The "phylogenetic footprint" is usually obtained by constructing a global multiple alignment of the orthologous promoter sequences using CLUSTAL_X (Thompson *et al.*, 1997) and then identifying conserved regions in the

alignment. However this approach does not always work as demonstrated by Cliften *et al.* (2001) for several *Saccharomyces* species. They discovered that if the species are too closely related, the sequence alignment is obvious but uninformative, because the functional elements are not sufficiently better conserved than the surrounding nonfunctional sequence. In contrast, if the species are too distantly related, it is difficult or impossible to find an accurate alignment.

Several algorithms have been implemented for comparative analysis such as "FootPrinter" which has successfully identified an element for the metallothionein gene family (Blanchette and Tompa, 2002). This approach has also been applied to conserved operons from 24 species to discover regulatory motifs using the AlignACE algorithm (McGuire and Church, 2000). Cross-species alignment between *C. elegans* and *C. briggsae* has been used to confirm the *cis*-acting element for the regulation of gene *dpy-7* (Gilleard *et al.*, 1997) and the heat shock elements detected by Guhathakurta *et al.* (2002a).

1.10: Thesis aims and objectives.

With the combined contributions of numerous researchers throughout the world, there are now in excess of several hundreds of published *C. elegans* gene expression patterns. An important next step is the interpretation of this data and the generation of meaningful conclusions with regard to gene regulation. With this in mind the primary aim of this Ph.D. project is to devise a computational strategy for the detection of *cis*-acting regulatory elements and thereby identify DNA sequences that control gene expression. The first stage in achieving this goal will be to perform a comprehensive analysis of *C. elegans* gene expression patterns to identify genes with similar expression patterns. Given that such genes are expressed at similar times and cell locations a number of these genes are likely to be co-regulated and contain *cis*-acting regulatory elements recognized by a common regulatory system. The co-regulated genes in question were identified using the reporter gene fusion approach (Section 3.2.1) in the laboratory of Dr I. A. Hope. In these studies a number of genes were identified which were either specifically or solely expressed in one particular cell type, namely the excretory cell [http://129.11.204.86:591]. The excretory cell is present as a single cell and is the largest cell in *C. elegans* (Sulston *et al.*, 1983). It is located on the ventral side of the organism in close proximity to the pharynx, and extends two processes

dorsolaterally over the ventral muscle quadrants to the lateral surface (Figure 1.5). The excretory cell is shaped like the letter 'H' and its primary function is in osmoregulation (Buechner, 2002). Given the close temporal and spatial expression of these genes it is probable that they are co-regulated and directed by the same regulatory systems. Thus, expression of this group of genes provides a useful data set for the *in silico* analysis of the promoter regions. Furthermore genes expressed in the excretory cell may be particularly useful for such an analysis because as there is only a single excretory cell per organism the mechanism of expression may potentially be simpler than in a cell type which is present in abundance.

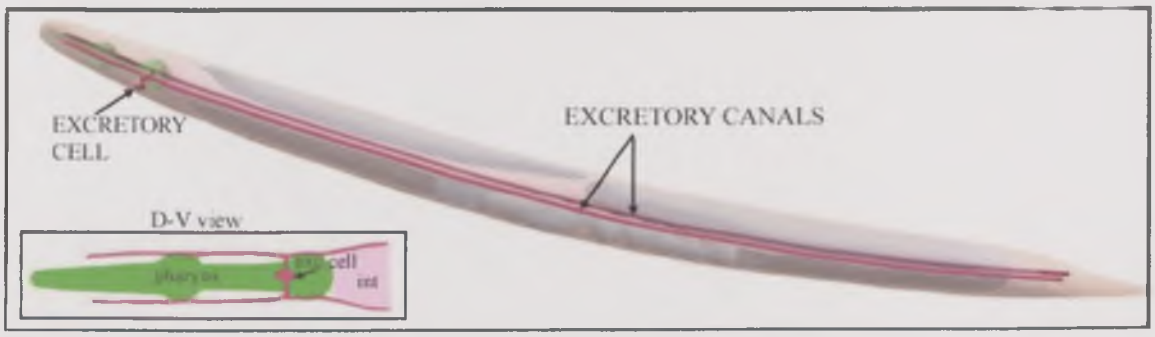

Figure 1.5. Location of the excretory cell in *C. elegans*. (Figure from wormatlas, [http://www.wormatlas.org]).

To identify *cis*-regulatory elements able to drive expression of these genes DNA sequences from the upstream region will be analyzed using two software algorithms, MEME (http://meme.sdsc.edu/meme/website/) and SPEXS (http://ep.ebi.ac.uk/EP/SPEXS/). Potential regulatory elements will then be assessed using a novel scoring strategy that incorporates biological factors to aid in the detection of potential candidates for *cis*-acting elements. It is anticipated that this comprehensive analysis will provide a valuable insight into the complex regulation of animal development.

1.11: Specific objectives of the Ph.D. research.

1. Survey the *C. elegans* scientific literature and extract gene expression pattern and expression profile data.

2. Convert gene expression data collated from the published literature into ace format and submit to WormBase/ACeDB.

3. Compare gene expression pattern data (generated by the use of two or more experimental methods (i.e., reporter gene fusion, mRNA *in situ* hybridization and immunostaining) in order to evaluate the consistency and reliability of the data obtained by these methods. In addition, compare and contrast expression profile data of temporal gene expression generated using northern blot assays, western blot assays, RT-PCR and microarray techniques.

4. Prediction of candidate *cis*-acting elements by analyzing the upstream DNA sequences of genes expressed specifically in the excretory cell. This will be achieved using:

- The MEME software.
- The SPEXS software.
- Assess the software output with a novel scoring strategy to identify the most likely *cis*-acting elements.
- Test the scoring strategy developed on a data set where *cis*-acting elements had previously been characterized.
- Test the motifs identified using the excretory cell data set on a data set consisting of genes showing expression in the excretory cell that were not used in any previous analyses.
- Test the motifs identified on a data set of *C.briggsae* orthologues of *C. elegans* genes that showed expression in the excretory cell.
- Search the *C. elegans* genome with the most likely candidate motif to predict genes that might show expression in the excretory cell.

1.12: Thesis composition.

The next chapter of this thesis, Chapter 2 contains details of the methods and procedures employed in this research project. The first chapter of results, Chapter 3

describes the procedure and importance of extracting gene expression patterns from the published literature and then assimilation into the WormBase/ACeDB database. In addition, it details the comprehensive comparison of gene expression pattern data characterized using the different experimental techniques available for *C. elegans*, to evaluate the consistency of data between the different approaches. Chapter 4 to 7 discuss the implementation of a strategy to identify potential *cis*-acting regulatory elements from a set of co-regulated genes. Chapter 4 includes an analysis of the promoter region of all the genes that show expression in the excretory cell using the MEME software. Chapter 5 contains the results from the SPEXS software analysis using the genomic DNA inserts from reporter gene fusion experiments, and also the development of a scoring strategy. Chapter 6 includes the results obtained from testing the scoring strategy on a known data set for which *cis*-acting elements have previously been characterized and the analysis of the 2 kb upstream region of the excretory cell data set. Chapter 7 contains the results obtained from *in silico* testing of the *cis*-acting regulatory element identified as possibly involved in the control of excretory cell expression and includes the results from searching the *C. elegans* genome with the most likely candidate *cis*-acting element in order to predict genes that may express in the excretory cell. The final chapter of this thesis will provide a general discussion of the research described in this Ph.D. thesis and directions for future work.

Chapter 2

Chapter 2: Materials and methods.

2.1: Surveying the scientific literature for *C. elegans* gene expression pattern information.

The search for publications containing *C. elegans* gene expression pattern information was initially performed in PubMed (http://www.ncbi.nlm.nih.gov/entrez) (Table 2.1). To reduce the number of articles returned, most of which simply referred to *C. elegans* rather than being about *C. elegans*, a search was performed in the ACeDB paper class which contains all published literature in the *C. elegans* field including *C. elegans* meeting abstracts and Wormbreeders gazette articles as well as full publications in peer-reviewed scientific journals. Many reports could be eliminated by reading the titles or abstracts with closer examination of the full text eliminating more reports. The remaining publications yielded expression pattern descriptions that were extracted using the expression pattern class model to form .ace files for each.

List of keywords	
Pubmed	***ACeDB "paper" class***
Expression AND *C. elegans* LacZ AND *C. elegans* In *situ* AND *C. elegans* Antibody AND *C. elegans* Immunostaining AND *C. elegans* Expression patterns AND *C. elegans* GFP AND *C. elegans*	Expression LacZ GFP Antibody In *situ* hybridization Immunostaining Promoter activity Expression + LacZ Expression + GFP Expression + antibody Expression + immunostaining Expression + in *situ* hybridization Reporter expression pattern Antibody + reporter + in *situ* Immunostaining + in *situ* + GFP Immunostaining + in *situ* + lacZ

Table 2.1. Keyword combinations used for searching the PubMed and the ACeDB databases for papers containing *C. elegans* gene expression patterns.

2.2: Extracting DNA sequences for analysis with MEME and SPEXS programs.

The methods described in this thesis have been implemented in C language for PCs running Linux (Suse, [http://www.suse.com]). The gene expression patterns used in the detection of *cis*-acting elements are available from the WormBase database (http://www.wormbase.org).

2.2.1: Extracting sequences of genomic DNA inserts for analysis by SPEXS.

Initially 18 gene expression patterns known to include expression in the excretory cell were utilized. These patterns had been determined using the reporter gene fusion approach in the laboratory of Dr. I. A. Hope, University of Leeds (Table 2.2). The entire sequence of the genomic DNA fragment, averaging approximately 5 kb in size, contained in the reporter gene fusions was used for analysis with the SPEXS software. These sequences constituted the positive set. A negative set was also utilized which composed of 33 genes that expressed in cell groups other than the excretory cell (Table 2.3). The negative data set was restricted to 33 insert fragments due to software limitation.

The cosmid sequences of all genes in both the positive and negative sets were downloaded from the NCBI Entrez website (http://www.ncbi.nlm.nih.gov). The search was performed under "Nucleotides" using the cosmid name as "keyword". When the cosmid sequences were downloaded, the genomic DNA insert sequences were extracted using a C program (command_line_seq_parse.c) (Figure 2.1). This C program (command_line_seq_parse.c) was implemented to read a FASTA format file and takes as input filename e.g. "Y18D10A.dna", genomic insert sequence start e.g. "88453" and sequence end e.g. "95131" as command line arguments (Figure 2.1). The output from this program (a genomic DNA insert sequence in FASTA format) was then redirected to the file with the name of the assayed gene e.g."Y17D10A.23" (Figure 2.1).

Gene	*Fragment endpoints*	*Cosmid/YAC name*	*Fragment size/bp*
B0285.6	partial *Hin*dIII 17581 - *Bam*HI 22836	B0285	5255
C14A4.12	*Avr*II 30802 - *Bgl*II 35777	C14A4	4975
C17H12.14	~7kb, partial *Sau*3A fragment from 13783	C17H12	7000
C46C2.1	partial *Hin*dIII 13202 – *Bam*HI 20225	C46C2	7023
F10B5.1	XbaI 1502 - *Bsp*EI 4838	F10B5	3336
F41E7.1	Partial *Hin*dIII (-590) - *Sal*I 5144	F41E7	5734
F44A2.5	partial *Hin*dIII 3036 - *Bam*HI 8925	F44A2	5889
F54C9.1	*Pst*I 10314 - *Bam*HI 5756	F54C9	4558
F54C9.7	*Pst*I 14314 - *Nhe*I 19241	F54C9	4927
R05G6.6	*Pst*I 31572 (cosmid F55G1) - *Bam*HI 4274	R05G6	7900
R12C12.6	partial *Hin*dIII 31417 - *Bam*HI 25464	R12C12	5953
R13H4.5	partial *Hin*dIII 24901 - *Bam*HI 30704	R13H4	5803
T14E8.1	partial *Hin*dIII 14032 - *Bam*HI 20323	T14E8	6291
T20B5.3	partial *Hin*dIII 27163 – *Nhe*I 19502	T20B5	7661
Y18D10A.23	Partial *Hin*dIII 95131 – *Sal*I 88453	Y18D10A	6678
Y62E10A.1	partial *Hin*dIII 15219 - *Bam*HI 11218	Y62E10A	4001
Y70G7B.3	partial *Hin*dIII 14442 - *Bam*HI 21077	Y70G10A	6635
**Y113G7B.24*	52259 - 47852	Y113G7B	4407

Table 2.2. Genes contained in the positive set used for the analysis of the full genomic DNA insert sequences. The fragment size and endpoints, mostly with respect to restriction enzyme site and cosmid/YAC sequence coordinates are provided. * Gene fusion constructed with PCR product.

Gene	*Fragment endpoints*	*Cosmid/YAC name*	*Fragment size/bp*
B0034.1	*Pst*I 24815 – *Sal*I 28362	B0034	3447
B0228.7	*Hin*dIII 38416 – *Bam*HI 33476	B0228	4940
B0280.1	*Pst*I 27387 – *Bgl*I 31972	B0280	4585
B0280.4	*Hin*dIII 4140 – *Bgl*I 7107	B0280	2967
B0464.4	*Xba*I 19864 – *Spe*I 16882	B0464	2982
B0495.9	*Hin*dIII 21331 – *Pst*I 17820	B0495	3511
B0495.10	*Xba*I 30470 – *Bgl*I 24254	B0495	6216
B0523.5	*Pst*I 10439 – *Bsp*EI 2434	B0523	8005
C09F9.3	Partial *Hin*dIII (- 34665) – *Bam*HI 28187	C09F9	6478
C17G10.1	*Hin*dIII 12743 – *Pst*I 15573	C17G10	2830
C29H12.6	*Pst*I 4843 – *Xba*I 11232	C29H12	6389
C40H1.6	*Pst*I 18509 – *Nhe*I 22793	C40H1	4284
C45G7.6	Partial *Hin*dIII 35155 – *Bam*HI 32466	C45G7	2689
C50B6.8	Partial *Hin*dIII 28889 – *Sal*I 23947	C50B6	4942
C55C3.5	Partial *Hin*dIII 27718 – *Sal*I 23662	C55C3	3771
F01F1.12	*Hin*dIII 37521 – *Bsp*EI 33315	F01F1	4206
F01F1.6	*Hin*dIII 11654 – *Age*I 15719	F01F1	4065
F10F2.1	*Hin*dIII 11689 – *Pst*I 3379	F10F2	8310
F10F2.4	*Xba*I 15132 – *Bgl*I 17290	F10F2	2158
F14F4.3	Partial *Hin*dIII 26599 – *Bam*HI 21198	F14F4	5401
F28F8.5	Partial *Hin*dIII 33996– *Bam*HI 27679	F28F8	6317
F32A11.5	Partial *Hin*dIII (-5614) – *Bam*HI 1681	F32A11	1881
F38A1.5	Partial *Hin*dIII 16237 - *Pst*I 22553	F38A1	6316
F41C3.5	*Hin*dIII 19766 – *Bsp*EI 23341	F41C3	3575
F45D11.14	Partial *Hin*dIII 11221 – *Nsi*I 7062	F45D11	4159
F47A4.2	Partial HindIII 22026 – *Bam*HI 16290	F47A4	5736
F52F12.6	Partial *Hin*dIII 25160 – *Sal*I 18809	F52F12	6351
F53H8.3	Partial *Hin*dIII (-740) – *Bam*HI 4733	F53H8	3993
F58A3.2	Partial *Hin*dIII 8319 – *Bam*HI 13458	F58A3	5139
R11A5.2	Partial *Hin*dIII 11221 – *Nsi*I 7062	R11A5	4159
R12E2.7	Partial *Hin*dIII 11522 – *Bam*HI 6740	R12E2	4782
T24C4.7	Partial *Hin*dIII 31320 – *Sal*I 25790	T24C4	5530
ZC123.1	Partial *Hin*dIII 30764 – *Nhe*I 37246	ZC123	6482

Table 2.3. Genes contained in the negative set for the analysis of full genomic DNA insert. The fragment size and endpoints, mostly with respect to restriction enzyme site and cosmid/YAC sequence coordinates are provided.

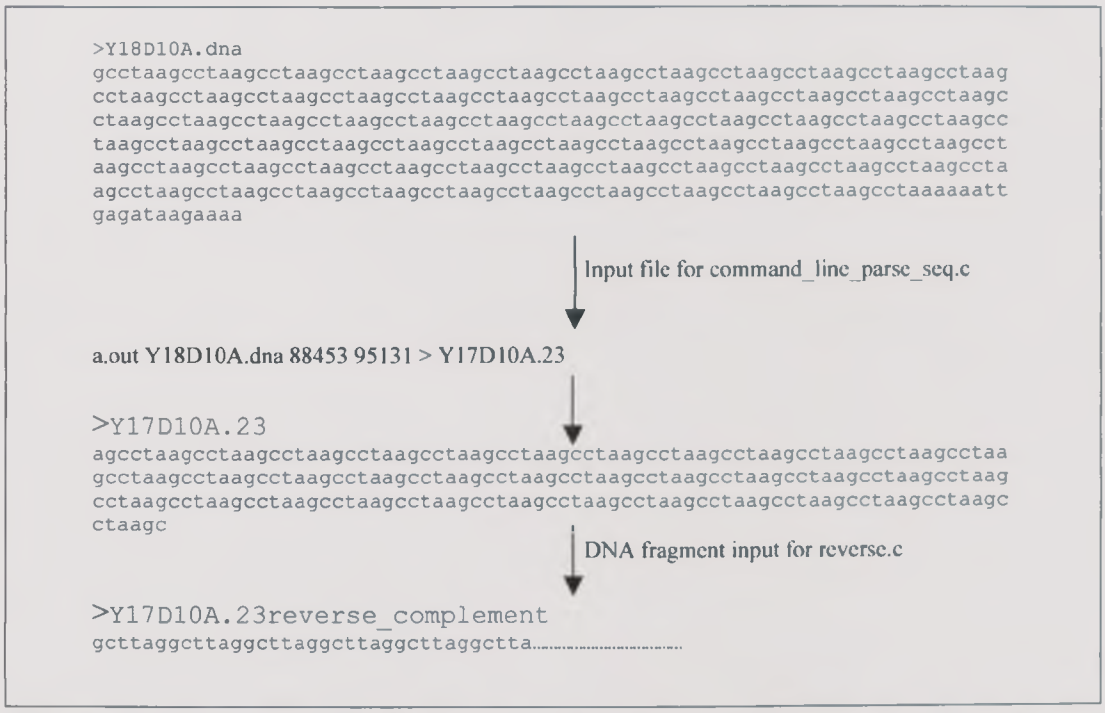

Figure 2.1. **The command line arguments used with the C program "command_line_seq_parse.c" to extract the genomic DNA insert sequences for all genes.**

2.2.2: Extracting DNA sequences from the upstream region of genes.

DNA sequences from the upstream regions of genes from the positive and negative sets were obtained in a similar manner to that described above (Section 2.2.1). Due to concurrent, on-going experimental work two additional genes that expressed in the excretory cell were characterised and subsequently included in the positive set, (details shown in Table 2.4). It was possible to include a greater number of genes in the negative set because of the reduction in the size of the DNA sequence being assayed (Table 2.5). Genes from the positive (n=20) and negative (n=134) sets were organised into six groups depending on the chromosome from which they were derived. The gene-start and gene-end values, as well as the gene orientation were obtained from GFF (General Feature Format, [http://www.sanger.ac.uk/Software/formats/GFF]) files, downloaded from the Wellcome Trust Sanger Institute website (http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/GFF_files.shtml). In addition the files containing the chromosomal DNA sequences and the chromosomal

exon and intron files were also downloaded. The DNA sequence files were used to extract the DNA sequences for the 2 kb upstream regions of each of the genes from the positive (Tables 2.4) and the negative (Tables 2.5) sets using the C program (command_line_seq_parse.c) as described above (Section 2.2.1). The examples below demonstrate how the 2 kb upstream regions of each of the genes were evaluated depending on their orientation (also provided in GFF file):

Examples

If gene was antisense orientated (3' - 5'):

Y18D10A.23 located on chromosome I, orientation (-), gene start 12767494 and gene end 12769955.

2 kb upstream of gene start would be:

x = gene_end + 2 kb

i.e. 12711955 = 12769955 + 2000

If gene was sense orientated (5' - 3')

C14A4.12 located on chromosome II, orientation (+), gene start 10612922, gene end 10618065.

2 kb upstream of gene start would be:

x = gene start- 2 kb

i.e. 10610922 = 10612922 –2000

2.2.3: Reverse complementation of DNA sequences.

The isolated DNA sequences were then interpreted by the program, "reverse.c", thereby producing a reverse complement (in FASTA format) of the sequence, which was then redirected to another file (Figure 2.1).

All the genomic insert DNA sequences extracted using this program were then verified by checking the sequence in ACeDB (local version). In ACeDB the search was performed under "sequences" using the cosmid name as "keyword". On the cosmid sequence map window a restriction digest was performed using the enzymes used in the original

experiment and the DNA sequence of the genomic insert was verified "by direct inspection".

Once verified, all genomic insert sequences from the positive set were concatenated into one file for submission to subsequent analysis. This was repeated with the negative data set.

Gene in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome I	
	Y18 D10A.23
Chromosome II	
C14A4.12	F54C9.1
F10B5.1	*R12C12.6*
F54C9.7	*T21B10.5*
	Y46G5A.4
Chromosome III	
B0285.6	
Y70G10A.3	
Chromosome IV	
C17H12.4	*Y62E10A.1*
C46C2.1	
R05G6.6	
Chromosome V	
F44A2.5	*Y113G7B.24*
R13H4.5	
Chromosome X	
F41E7.1	*T20B5.3*
T14E8.1	

Table 2.4. Chromosomal assignment of genes contained in the 2 kb excretory positive set.

Genes in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome I	
B0207.3	*F52F12.6*
C30F8.4	*R11A5.2*
C48B6.2	*R12E2.7*
F40E3.2	*T08G11.1*
T10B11.3	*W05F2.4*
ZC123.1	
ZC308.2	
Chromosome II	
B0034.1	*B0228.7*
C17G10.1	*B0495.9*
C29H12.1	*B0495.10*
F32A11.5	*C09F9.3*
F35D11.2	*C17G10.5*
F41C3.2	*C25H3.11*
F41C3.5	*C29H12.6*
F41C3.8	*D2085.5*
F45D11.14	*F07A11.6*
F54C9.5	*F46C5.6*
R05F9.1	*F52H3.3*
W01G7.4	*F54C9.11*
W02B12.11	*F54D5.1*
Y48C3A.16	*T01H3.3*
Chromosome III	
B0280.1	*B0464.4*
B0280.4	*B0523.5*
C07G2.1	*C05B5.3*
C36A4.9	*C29E4.5*
C40H1.6	*C46F11.5*
F01F1.2	*D2007.5*
F01F1.5	*F01F1.10*
F01F1.6	*F01F1.12*
F02A9.3	*F02A9.4*
F10F2.4	*F10F2.1*
F42A10.2	*F10F2.2*
F42H10.9	*F42A10.3*
F44B9.2	*F54C8.4*
H10E21.1	*F54F2.8*
K08E3.3	*T12A2.15*
K08E3.8	*T24C4.7*
M03C11.3	*Y47D3A.2*
R08D7.3	*Y76A2B.5*
Y47D3A.16	
Y75B8A.27	
ZC21.3	
ZK637.8	
ZK643.3	

Table 2.5. **Chromosomal assignment of genes contained in the 2 kb excretory negative set.**

Genes in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome IV	
C25A8.4	*C45G7.5*
C49C3.5	*C45G7.6*
F47C12.2	*C55C3.5*
F55G1.6	*F38A1.5*
R11A8.6	*M70.5*
W03D2.1	*T04B2.5*
W03F8.4	*T09A12.4*
Y73B6BL.9	*Y37A1B.14*
Y105C5A.15	
Chromosome V	
F47B8.6	*B0024.10*
K09H11.3	*C04E12.7*
T01G6.2	*C50B6.8*
T25E12.4	*F07C3.4*
W05B10.4	*F28F8.5*
	R11D1.11
	T26H2.9
	Y40B10A.9
Chromosome X	
C02B4.1	*C33D12.5/6*
C36C9.1	*C44C10.1*
C39D10.3	*C49F8.2*
C46C11.2	*F14F4.3*
F14B8.3	*F40E10.6*
F18G5.2	*F47A4.2*
F52E4.1a	*F53B3.3*
F53H4.5	*R09A8.5*
F53H8.3	*T27A8.2*
F58A3.2	
H22K11.1	
M03B6.2	
T28B4.1a	
ZK813.3	

Table 2.5 continued. Chromosomal assignment of genes contained in the 2 kb excretory negative set.

2.3: MEME analysis.

The source code of the MEME software version 2.2 for the LINUX operating system was downloaded and installed from the San Diego Supercomputer Centre ftp site (ftp://ftp.sdsc.edu/pub/sdsc/biology/meme/old_versions/). The MEME software requires input DNA sequences to be in the FASTA format (as shown below, Figure 2.2).

```
>B0304.1
cttatttcaggaaaatttttttcaaaactgtaaaacaaaaaccatttttcacagaatctaa
agggtatctgaaagcttaaaataacttcagaaagatatcaattccagctgtttagtacct
gaactgtctgtaaacgtttcttctcgaattatagaaaatttttccacttttttcaagttcag
>B0304.1reverse_complement
ttctggaaaattattggaaaatttggaaaatggtagaaatagaaaaaatattaaagaata
ttaataagttgagacaaagtaaaacgtgtttttttttttgaaaaatagcatatatagcga
aggttaggttctactttgataaccgacatttcgattcccttatattttaacaaacaa
```

Figure 2.2. An example of FASTA format file. The first line is a comment line beginning with ">" followed by gene name and then on the next line the DNA sequence. All lines beginning with ">" have the ">" removed so that only the DNA sequences are used in the analysis.

After concatenation of the DNA sequences into a single file (Table 2.6, Section 2.2.3), MEME was executed with the following command line arguments:

meme /home/archana/sequences/seqs/posseqs.seq -dna -mod tcm -nmotifs 10 -W 20

The command line argument "meme" was followed by the path for the file containing the DNA sequences "/home/archana/sequences/seqs/posseqs.seq". The options "-dna" indicates which alphabet to use in the program (i.e. ATCG rather than the amino acid codes). This was followed by another option (-mod) to indicate the number of times the motif was expected to occur within a sequence. There are three options under this category, OOPS (one occurrence per sequence), ZOOPS (zero or one occurrence per sequence) and TCM (any number of occurrences per sequence). The other options used were –nmotifs and –W; the former refers to the number of motifs searched for to output and the latter indicates the desired length of the motif. The other option used in this analysis was the –neg option. The –neg option allows MEME to be executed with another file containing DNA sequences as a negative set (i.e. that lacks the element(s) being sought). The command "–neg" was followed by the path for the file containing the negative set of sequences. This option allows MEME to search the negative set for each of the motifs identified in the positive set.

Excretory Positive set	*Size of the intergenic regions immediately upstream from the translational start (bp)*
C14A4.12	12577
C17H12.14	13680
C46C2.1	7808
F10B5.1	2736
F41E7.1	2532
F44A2.5	5000
F54C9.1	3028
F54C9.7	535
R05G6.6	8366
R12C12.6	7288
T14E8.1	9333
T20B5.3	2892
Y18D10A.23	5000
Y62E10A.1	10202
Y70G10A.3	3449
Y113G7B.24	2133

Table 2.6. The 16 excretory cell expressing genes used for analysis with MEME.

2.4: SPEXS analysis.

The SPEXS software identifies motifs that are common to the input sequences. The SPEXS program binary for Linux and the manual were downloaded from the EBI website (http://industry.ebi.ac.uk/~vilo/SPEXS/, http://industry.ebi.ac.uk/~vilo/SPEXS/Manual). It takes as input a comment line beginning with # followed by the gene name and then on the next line the DNA sequence as a single string (contains no line breaks) (Figure 2.3). All lines beginning with "#" are removed by the program so that only the DNA sequences are used to identify motifs.

```
#B0304.1
cttatttcaggaaaatttttcaaaactgtaaaacaaaaaccatttttcacagaatctaaagggtatctga
aagcttaaaataacttcagaaagatatcaattccagctgtttagtacctgaactgtctgtaaacgtttctt
ctcgaattatagaaaatttccacttttttcaagttcag
#B0304.1reverse_complement
ttctggaaaattattggaaaatttggaaaatggtagaaatagaaaaaatattaaagaatattaataagttg
agacaaagtaaaacgtgttttttttttgaaaaatagcatatatagcga
```

Figure 2.3. An example of the SPEXS input file format.

The command line arguments used to execute SPEXS were "spexs -f positive_filename -f negative_filename". The output from this program was redirected to a file using command line arguments "spexs -f positive_insert.data -f negative_insert.data > spexs_insert.output" (Figure 2.4). The output file called "spexs_insert.output" was parsed by a C program (p_s_commandline.c). This program takes as input the SPEXS output file and extracts all motifs that satisfy the threshold criteria (motif length 20 bp).

```
# Start the creation of patterns:

# Set_of_Sets: Domain size is: 616243 and nr. of set-collections is 308
<empty> 308/616242 1:40/79972 2:268/536270
a 308/200569 1:40/26648 2:268/173921
c 308/107398 1:40/13318 2:268/94080
g 308/107398 1:40/13318 2:268/94080
t 308/200569 1:40/26648 2:268/173921
aa 308/84088 1:40/11330 2:268/72758
ac 308/30242 1:40/3762 2:268/26480
ag 308/31127 1:40/3866 2:268/27261
at 308/54958 1:40/7666 2:268/47292
.
.
```

at	308	54958	1	40	7666	2	268	47292

```
.
.
ttttttttgcaaaaaatgt 1/1 2:1/1
ttttttttttgcaaaaaaaa 2/2 1:1/1 2:1/1
ttttttttttttttttggc 1/1 2:1/1
ttttttttttttttttggt 1/1 2:1/1
ttttttttttttttttgg 2/2 2:2/2
```

Figure 2.4. An example of the SPEXS output file format. The output from SPEXS can be divided into nine columns. The first column contains bases constituting a motif (at). The second column refers to the number of sequences (308) containing the motif from both the positive (40) and negative (268) sets. The third column indicates the Total Frequency (includes multiple occurrences per fragment of DNA) of the motif in both positive (666) and negative (47292) sets. The fourth column indicates the file number (1 = first file, [positive set]). The fifth column denotes the frequency per fragment in the first (positive) file with the sixth column referring to the total frequency in that file. The seventh column (if present) indicates the second file (negative set) from which data for columns eight and nine were obtained. Finally, columns eight and nine denote the frequency per fragment and the total frequency in the negative set, respectively.

2.4.1: Frequency cut-off threshold values for the genomic insert approach.

As the SPEXS output was large, a threshold frequency was defined to reduce the number of motifs considered for further analyses. For the genomic insert analyses, initially it was calculated that a motif of width 8 bp would only randomly occur once in every 65 kb. The negative set consisted of approximately 190 kb DNA sequence and therefore an 8 bp motif is expected to occur with a frequency of three, as shown in the example below.

Example

There are four bases ATGC in DNA sequences. Thus, the chance occurrence of each base would be 1 in 4. Therefore the probability of a motif of 8 bp occurring by chance would be: $1/4^8 = 1/65536$

i.e. once in every 65 kb of DNA sequence.

On average the number of times an 8 bp motif would occurr by chance in 190 kb DNA sequence would be:

190/65 = 2.92 or approximately 3.

To potentially increase specificity a frequency threshold was decided upon which contained three more occurrences in the positive set than in the negative set (Section 5.2.1). All the motifs that satisfied the frequency cut-off value were then compiled in an *excel* (Microsoft, USA) software spreadsheet. All duplicate motifs (identical motifs other than opposite orientation) were identified by a C program (that performs an "all versus all" search with each motif within the list, [Figure 2.5]) and then discarded from the list. In addition, this program also identified motifs that were identical except for a mismatch of one base pair at an internal position or extended by a single nucleotide.

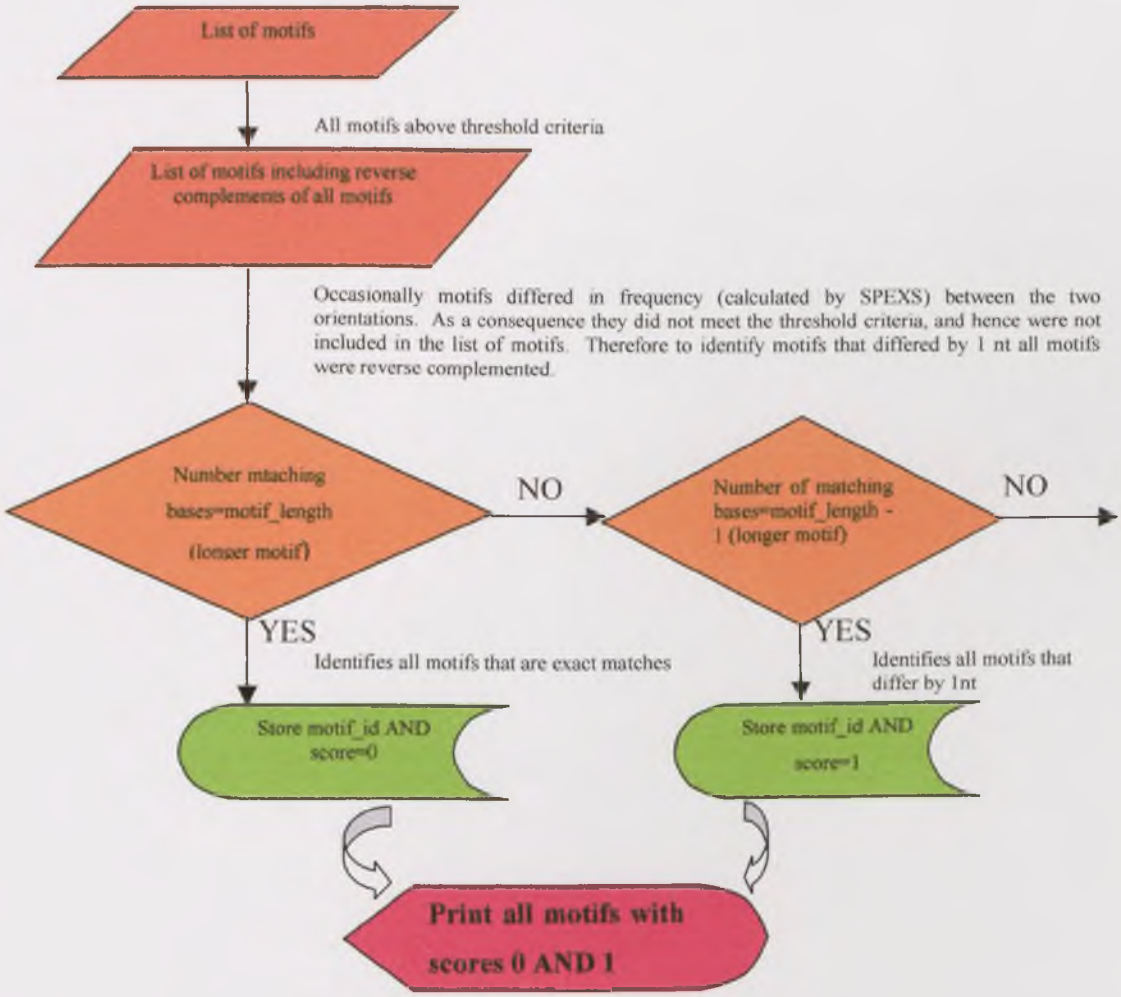

Figure 2.5. A flowchart of the C program implemented to identify all duplicated motifs and those that differ by a single nucleotide.

2.5: Obtaining frequency values of motifs from the positive and negative sets.

A program written in C language was implemented to search for each of the motifs, in a file containing the DNA sequences (i.e. the positive and negative sets). The program was designed to search for a string (i.e. a motif) within a larger string (i.e. DNA sequences). The number of occurrences of the motif within the DNA sequences (both positive and negative sets) was determined. The position within the sequence and the name of the sequence was then recorded and a print out on screen was generated which was then

redirected to another file (Figure 2.6).

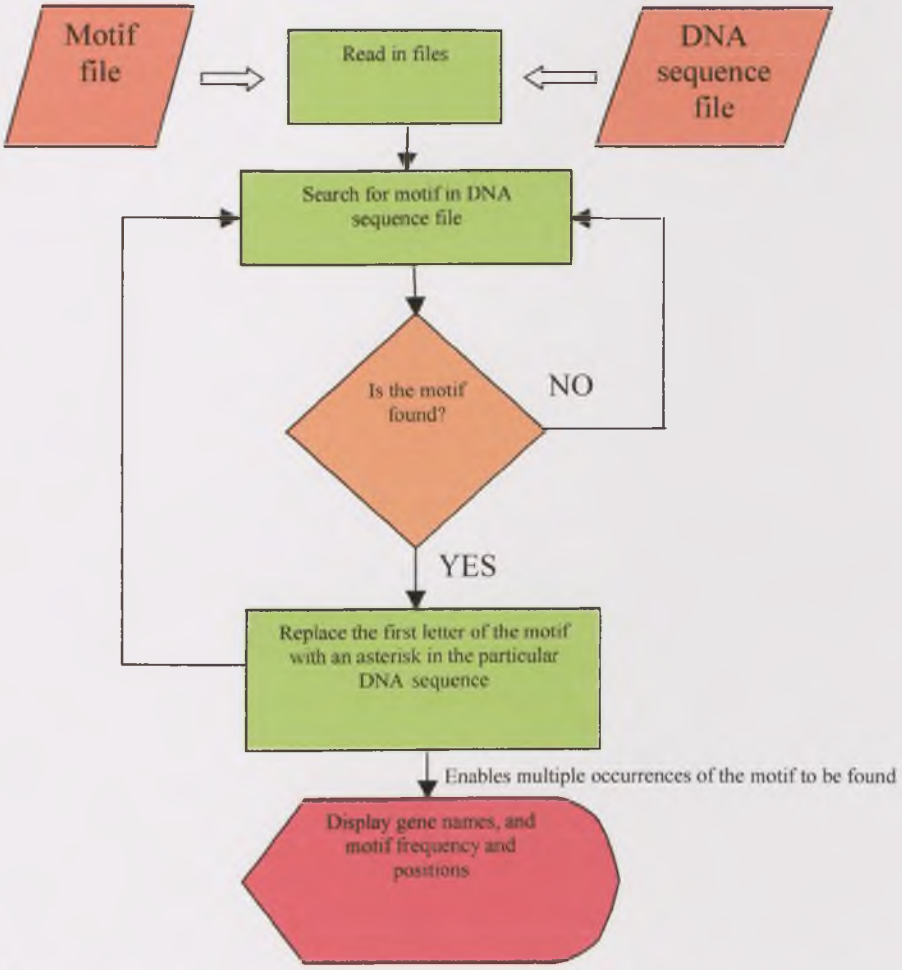

Figure 2.6. **A flowchart illustrating the process of obtaining frequency values for a particular motif in a DNA sequence file.**

The frequency values of the motifs from both the positive and the negative data sets were then entered into an *excel* spreadsheet (for example see Appendix II). The first column of this worksheet contains the list of motifs with columns 2 to 9 divided into two distinct categories. Columns 2 to 5 consist of frequency values (fragment frequency, [the number of sequences with one or more occurrences] and total frequency, [the total number of all occurrences in all sequences]) obtained from the positive set for each of the motifs in both

forward and reverse orientations. Columns 5 to 9 consist of the frequency values obtained from the negative set.

2.6: Determining the position of motifs contained in the DNA sequences of the positive data set.

2.6.1: Genomic insert approach.

A C language program was implemented to identify the location of each occurrence of a motif in terms of intergenic, intron or exon region. Motifs in the intergenic region were further categorised into greater than 1kb, or less than 1 kb, upstream of the initiation codon.

Although the location of each motif in the genomic DNA insert fragment was known it was difficult to stipulate the exon, intron and upstream structure for each of the insert fragments because of variations in positions in each of the genomic DNA insert fragments. Therefore to overcome this problem a system call was made from within the program to the FASTA (Pearson and Lipman, 1988) algorithm using the genomic insert DNA sequence as query to search the specific chromosomal DNA sequence. This search resulted in an alignment of the genomic DNA insert with the matching chromosomal DNA sequence contained in a FASTA output file. The FASTA output file was then parsed to extract the chromosomal positions of the start and the end values of the genomic DNA insert alignment on the chromosome.

Once the start and end points of the genomic DNA inserts were determined within the respective chromosome DNA sequence, the location of the motif within the chromosome could then be determined. If the genomic DNA insert fragment was in the sense strand then the position value of the motif within the genomic DNA insert fragment was added to the start value of the genomic DNA insert fragment within the chromosome. Conversely, if the genomic DNA insert fragment was in the antisense strand then the end position value of the motif within the genomic DNA insert fragment was subtracted from the end value of the genomic DNA insert fragment within the chromosome.

After the position of each motif within the chromosome was determined, this value was

then used to compare the gene start and end values of all the genes predicted within the particular chromosome GFF "genes" file. If the chromosomal motif position value was greater than the gene start value and less than the gene end value then this procedure was repeated first with the particular chromosome GFF "intron" file and then with the GFF "exon" file. This classified the location of the motif to either intron or exon. However, if the chromosomal motif position value was lower than the gene start (for genes in sense strand) then it was categorised as in an intergenic region. An overview of this program is represented in a flowchart shown in, Figure 2.7.

2.6.2: Determining position of motifs within the 2 kb upstream region.

Motifs that were located within 2 kb of the region upstream from the translational start were further divided into two categories; as described in Section 2.6.1. This precise positioning was achieved for all motifs in the positive set using the C program "position.c". The resultant two data sets were then searched with each of the motifs. The frequency values obtained using each of the files were entered into the $\geq$ 1 kb and $\leq$ 1 kb columns in the *excel* file. The sum of the values from these two columns was verified against the total frequency of occurrence values (obtained in Section 2.5). Any discrepancies were examined "by direct inspection". Discrepancies occurred rarely and were due to the motifs overlapping the boundary of the 1kb region.

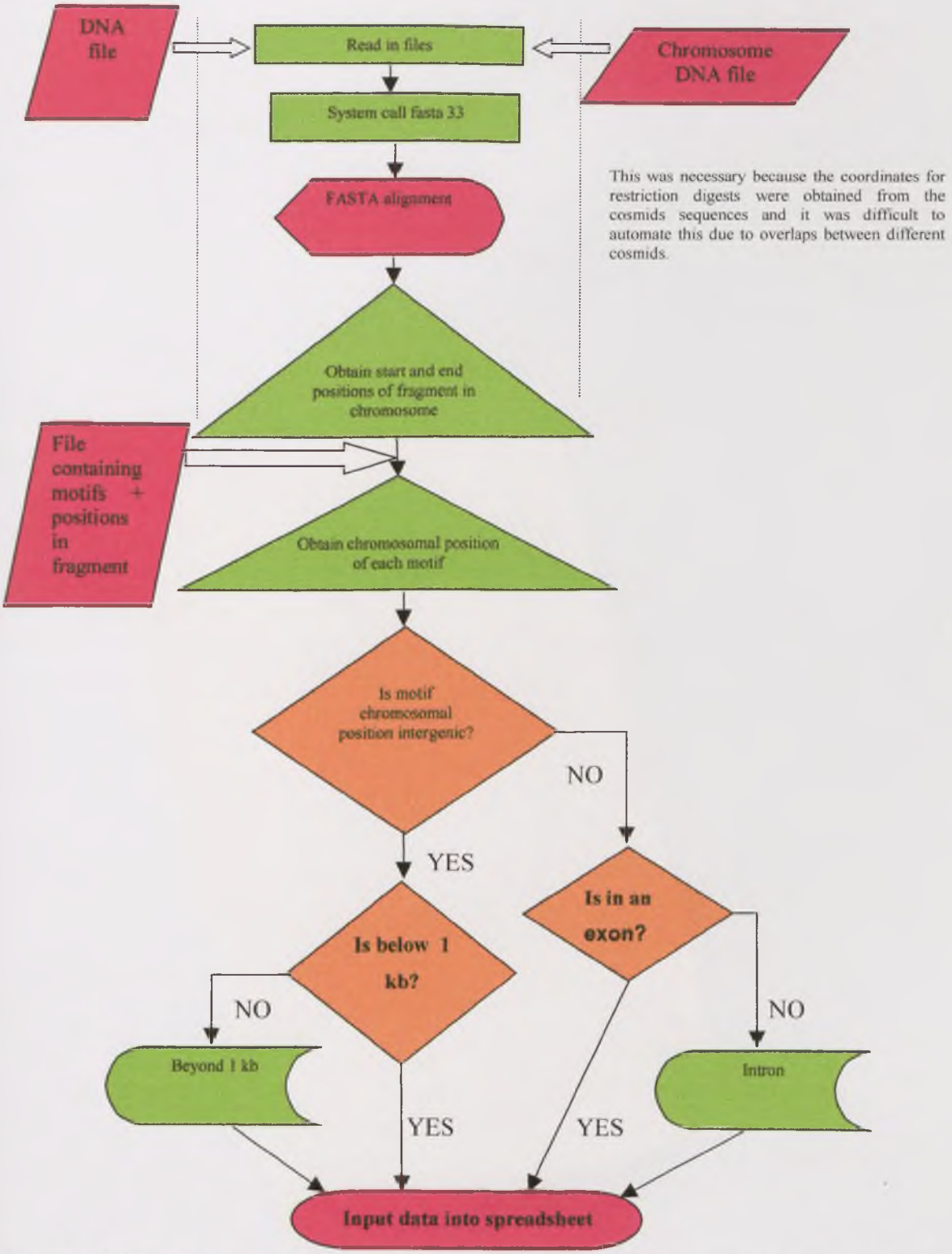

Figure 2.7. **A flowchart illustrating the steps involved in defining the location of each motif in the genomic DNA insert analysis.**

2.7: Determining the DNA sequence complexity of each of the motifs.

The DNA sequence complexity value for each of the motifs was determined using Equation 1 originally defined by Wootton *et al.* (1993) for amino acid sequences but adapted for DNA sequences.

Equation 1

$$Complexity = -\sum_{i=1}^{4} Pi \ln Pi$$

Complexity = - (P1 ln P1 + P2 ln P2 + P3 ln P3+ P4 ln P4)
Pi = proportion of bases (AGTC) i

Examples

DNA sequence, AAAAAAAAAA

P1 (A) = 10/10 = 1
P2 (G) = 0
P3 (T) = 0
P4 (C) = 0

Complexity = - (1 ln 1 + 0 + 0 + 0)
= 0

DNA sequence, AGAGAGAGAG
P1 = 5/10
P2 = 5/10
Complexity = - (0.5 ln 0.5 + 0.5 ln 0.5)
= 0.69

2.8: Testing of scoring strategy with previously defined *cis*-acting element.

The scoring strategy to predict *cis*-acting elements was tested on data in which a motif had previously been identified from a group of potentially co-regulated genes. The motif identified was cccgCGGGagcccg, a *cis*-acting element involved in the expression of genes in the muscle cell group (Guhathakurta *et al.*, 2002b). Guhathakurta *et al.* (2002b) identified this muscle cell-derived motif using 2 kb of DNA sequence upstream from the

translational start. In this earlier study the positive set consisted of 19 genes, and the negative set consisted of 3000 randomly selected genes. To perform an analysis as closely comparable as possible to the study of Guhathakurta *et al.* (2002b), in the current analysis the positive set consisted of 2 kb DNA sequences upstream from the translational start of the same 19 genes (Table 2.7). However (unlike Guhathakurta and colleagues), the negative set consisted of 90 genes (Table 2.8) that were not randomly selected, but were from the Expression Pattern Database, Leeds University on the basis that they did not express in the following cell groups; muscle, pharynx, sphincter, and gonad. This selection criterion was performed to exclude all genes that expressed in the muscle cells and therefore potentially shared a similar mechanism of expression.

The 2 kb DNA sequences from the upstream region of genes from the positive and negative sets were extracted as described above (Section 2.2.2). Two files containing the 2 kb of DNA sequence of genes (in both orientations) in the positive set and negative set were used to execute SPEXS using the same command line options as in Section 2.4.

The values for frequency, position and DNA sequence complexity were obtained as described in, Sections 2.5, 2.6.2 and 2.7, respectively. These values were used to calculate the individual scores for each of the factors, which were then summed to obtain the Total Scores for each of the motifs. All the steps involved in this analysis are illustrated in Figure 2.8.

Genes in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome I	
C09D1.1	*F07A5.7*
D1081.2	*F11C3.3*
W10D5.1	*F54C1.7*
	Y105E8B.1c
Chromosome II	
B0304.1	*ZC101.2b*
Chromosome III	
F09F7.2	
F30H5.1	
Chromosome IV	
	ZK617.1a
Chromosome V	
T04C12.6	*K12F2.1*
	T04C12.4
	T04C12.5
Chromosome X	
C36E6.3	*F42E11.4*
M03F4.2	

Table 2.7. **Genes contained in the muscle positive set.**

Genes in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome I	
C10H11.1	*F28D9.1*
F09C3.1	*ZK524.3*
F26A3.5	
K07A1.2	
R06C7.8	
T02E1.8/F08A10.1	
ZC123.1	
ZK973.1	
Chromosome II	
B0034.1	*B0228.7*
C17G10.1	*B0495.10*
C29H12.3	*F10E7.9*
C32B5.6	*K02C4.4*
C50E10.10	
F15A4.3	
F32A11.5	
F40F8.8	
F41C3.3	
F41C3.5	
F45D11.14	
T16A1.1	
Chromosome III	
B0280.1	*B0336.7*
B0280.4	*B0464.4*
B0285.6	*F01F1.12*
C29E4.3	*F01F1.9*
C38C10.1	*F10F2.1*
C40H1.6	*F22B7.9*
C45G9.2	*F23H11.8*
F01F1.6	*F26A1.8*
F10F2.4	*F44B9.5*
F43C1.1	*F54F2.7*
F54G8.2	*F59B2.13*
K02D10.1	*R01H10.6*
R08D7.5	*Y47D3A.2*
T23G5.5	*ZC84.3*
Y70G10A.3	*ZK637.13*
Y75B8A.4	*ZK643.5*
Chromosome IV	
C25G4.10	*C17H12.6*
C46C2.1	*C33H5.12*
F55G1.6	*C45G7.6*
Y11D7A.8	*C252D10.11*
Y116A8C.33	*C55C3.5*
	K08C7.4
	T22B11.5
	Y61A9LA.1
Chromosome V	
C05E4.1	*C18C4.9*
H12C20.3	*C50B6.8*
R13H4.5	*F10D2.4*
T13F3.3	*K12B6.4*
T22H9.4	
Chromosome X	
C11E4.6	*C29F7.3*
C28G1.1	*C42D8.8*
F38B6.4	*R09G11.2*
F53H8.3	*T04G9.1*
ZK455.1	*T21H8.1*

Table 2.8. Chromosomal assignment of genes contained in the muscle negative set.

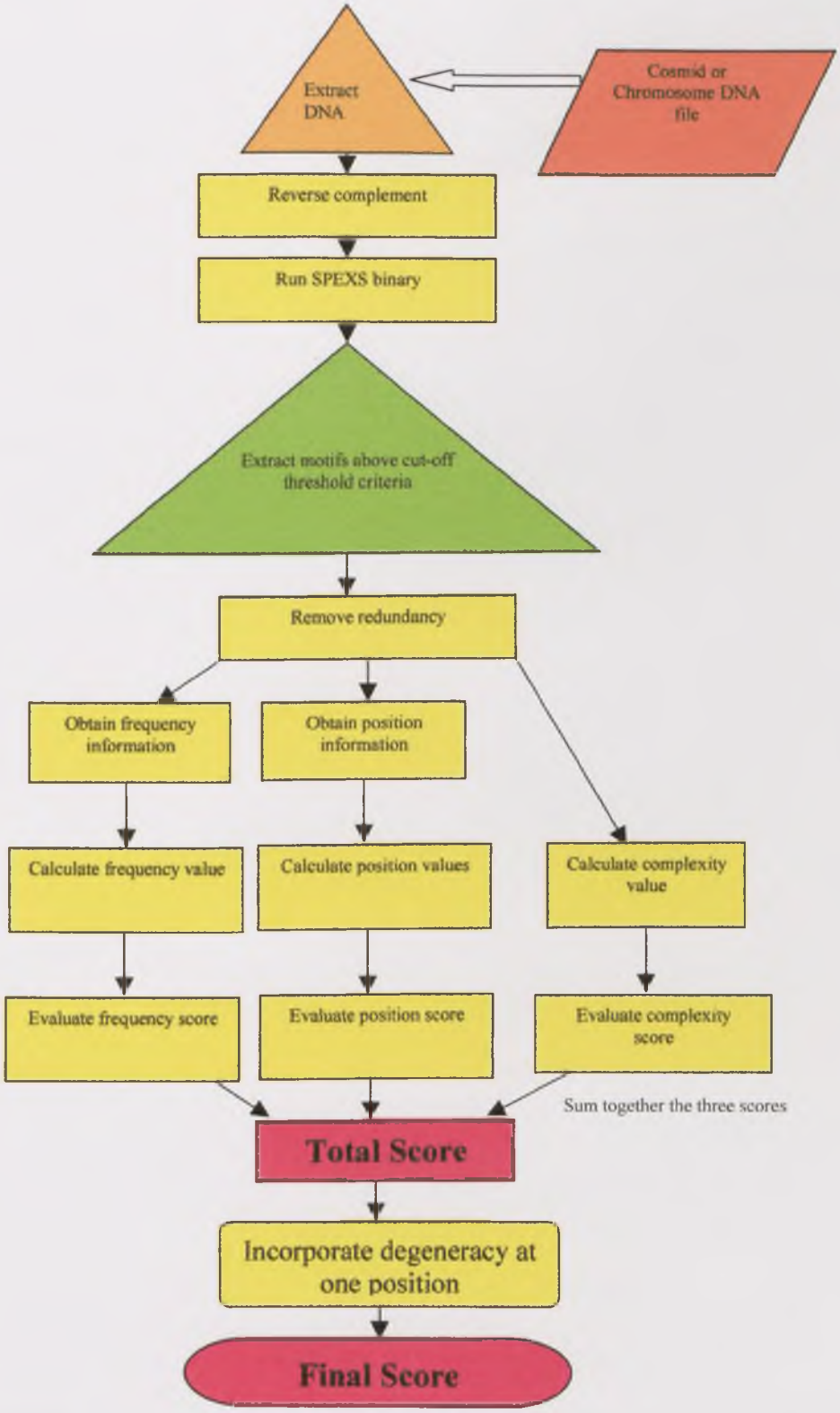

Figure 2.8. **A flowchart illustrating the steps involved in the evaluation of the Final Score of each motif detected by the SPEXS software.**

2.9: Testing in excretory cell-expressing genes.

A number of genes although expressing in the excretory cell, were not used in the earlier analyses, but were rather set aside to use as a test gene set (Section 7.2). The excretory test set consisted of 23 genes (Table 2.9) which were identified by searching within WormBase using the keyword "excretory cell" within the "Expression Pattern" class using the search tool found under "Advanced search" on the WormBase website (http://www.wormbase.org). This search also identified four other genes (*cdh-3, inx-13, lin-60* and *vha-4*) for which the cosmid genes names were undetermined and therefore the sequence information required for the analysis could not be obtained.

The negative set was identical to that used in the 2 kb excretory cell analysis (Table 2.5). The 2 kb DNA sequences from the upstream regions of the genes of the excretory test set were obtained in the same manner as described in Section 2.2.2. The frequency of occurrence for each of the motifs in the 2 kb excretory set were obtained as outlined in the flowchart, Figure 2.6.

Genes in forward orientation (+)	*Genes in reverse orientation (-)*
Chromosome I	
K04G2.8/apr-1	*K02B12.1/ceh-6*
R06A10.2/gsa-1	*K03D10.1/kal-1*
	T08G11.2/egl-32
	F55C7.7/unc-73
Chromosome II	
	E04F6.11/clh-3
	F45E10.1/unc-53
Chromosome III	
T17E9.1/kin-18	*R107.8/lin-12*
R10E11.8/vha-1	
R10E11.2/vha-2	
Chromosome IV	
F33D4.2A/itr-1	*C33D9.1/exc-5*
F35H10.4/vha-5	*Y38F2AL.3/vha-11*
Chromosome V	
Y75B12B.5/cyp-3	*R12G8.2*
F56H9.5/lin-25	
R31.1/sma-7	
Chromosome X	
ZK455.7/pgp-3	*C09B8.7/pak-1*
	T06F4.2/clh-4

Table 2.9. **Genes contained in the excretory cell positive test set.**

2.10: Identification of *C. briggsae* orthologues.

Translated protein sequences were obtained from WormBase for each of the genes expressing in the excretory cell of *C. elegans* (Table 2.4 and Table 2.9). This protein sequence information was then used to search the *C. briggsae* genome using the BLASTX algorithm (Altschul *et al.*, 1990). The BLASTX algorithm searches for the presence of a protein sequence in a database of translated DNA sequences. Genes were only considered strong orthologues when there was good overall alignment (including the start methionine) between the two genes. Potential orthologues were confirmed by searching for the closest homologue of the *C. briggsae* gene in the *C. elegans* genome using the BLASTX algorithm; orthology was confirmed if this identified the same *C. elegans* gene as had been used in the initial search of the *C. briggsae* genome.

Following the identification of *C. briggsae* orthologues, the respective translational start position and the orientation of each of the orthologues were determined. The 2 kb DNA sequence from the upstream regions of each of the genes were extracted as with previous analyses (Section 2.2.2). In addition the *C. briggsae* cosmids containing the particular orthologues were downloaded from the Wellcome Trust Sanger Institute web site (http://www.sanger.ac.uk/Projects/C_briggsae).

The *C. briggsae* orthologues (comprising the "*C. briggsae* test set") consisted of 26 genes (Table 2.10). The frequency of occurrence of each of the motifs in the 2 kb excretory set were obtained in a similar manner to that described in flowchart, Figure 2.6.

Genes in forward orientation (+)	***Genes in reverse orientation (-)***
F41E7.1	*C17H12.14*
R05G6.6	*F10B5.1*
T14E8.1a	*F54C9.1*
**Y46G5A.4*	*R13H4.5*
Y70G10A.3	*Y18D10A.23*
Y113G7B.24	*Y62E10A.1*
K04G2.8a/apr-1	*E04F6.11a/clh-3*
F33D4.2a/itr-1	*Y75B12B.5/cyp-3*
T17E9.1a/kin-18	*T08G11.2/egl-32*
R31.1/sma-1	*R06A10.2/gsa-1*
R10E11.2/vha-2	*K03D10.1/kal-1*
F35H10.4/vha-5	*C09B8.7a/pak-1*
Y38F2AL.3/vha-11	*ZK455.7/pgp-3*

Table 2.10. ***C. elegans*** **gene orthologues of genes contained in the *C. briggsae* positive test set.** * Gap in intergenic sequence beyond 1 kb.

Chapter 3

Chapter 3: A comprehensive analysis of *C. elegans* gene expression data generated using different experimental techniques.

3.1: Introduction.

The *C. elegans* genome sequencing project was largely completed in 1998. Based on the data generated, the *C. elegans* sequencing consortium (1998) predicted that there are approximately 19, 000 genes in the genome. Since the completion of the sequencing project research has been redirected to the functional characterization of these predicted genes. One key approach in this goal is the cataloguing of gene expression patterns. A gene expression pattern is a description of the location and timing of the expression of a particular gene. Tight regulation of the spatial and temporal synthesis of a gene product is vital for the function of all multicellular organisms during embryonic development and beyond. Control over the synthesis of a gene product may be exerted at multiple steps in the process. This control may influence transcription, the processing, stability and translation of mRNA or protein turnover. *C. elegans* is particularly well-suited to the generation of expression patterns because the largely-invariant developmental cell lineage has been completely determined, from zygote to adult, and therefore gene expression patterns can be described precisely at the cellular level.

The key repository for *C. elegans* gene expression pattern information is WormBase/ACeDB (for database URLs see Table 3.1). WormBase/ACeDB includes expression patterns provided by several research groups as well as many patterns extracted from published literature (Hope *et al.*, 1996; Stein *et al.*, 2001). The data consists of text descriptions, links to images, experimental details, references and authors. Proteome is a text-based database that also contains published expression pattern descriptions for *C. elegans*. The NEXTDB database contains gene expression pattern data generated in Yuji Kohara's laboratory (Mishima, Japan), by mRNA *in situ* hybridization, which has not been included in WormBase. These data consist of expression pattern images for thousands of genes, from zygote to adult, but with little annotation.

Gene expression pattern data are an important aspect of genome sequence interpretation and expression pattern databases have been established for many species (Table 3.1). FlyBase contains *Drosophila melanogaster* gene expression pattern data organised by tissue of expression (Flybase Consortium, 1998). The Berkley Drosophila Genome Project database has searchable, mRNA *in situ* hybridization data, with images, for approximately 1000 genes. FlyView has more than 3700 expression pattern images for enhancer trap lines, with associated text descriptions (Janning, 1997). There is a mouse genome informatics database which contains searchable mouse gene expression pattern data (Ringwald *et al.*, 2000; Ringwald *et al.*, 2001). In addition, a digital atlas of mouse development is currently being developed as a basis for presentation of the mouse gene expression pattern data (Davidson *et al.*, 2001). Bodymap contains data on human and mouse genes expressed in various tissues or cell types at various stages (Hishiki *et al.*, 2000). There is also a *Xenopus laevis* expression database with a structure similar to that of WormBase/ACeDB (Pollet *et al.*, 2000).

Databases	***URLs***
WormBase	http://www.wormbase.org
Proteome	http://www.proteome.com/DB-demo/intro-to-WormPD.html
NEXTDB	http://nematode.lab.nig.ac.jp
Xenopus	http://www.dkfz-heidelberg.de/abt0135/axeldb.htm
Bodymap	http://bodymap.ims.u-tokyo.ac.jp
Mouse	http://www.informatics.jax.org/menus/expression_menu.shtml
Flybase	http://flybase.bio.indiana.edu
Berkeley Drosophila GP	http://www.fruitfly.org/EST
FlyView	http://pbio07.uni-muenster.de

Table 3.1. Expression pattern databases available for different organisms with their URLs.

3.2: Approaches used to determine gene expression patterns in *C. elegans*.

There are three main approaches used for the determination of *C. elegans* gene expression patterns, each measuring a different aspect of gene expression. Reporter gene fusion approaches primarily assess promoter activity (Hope *et al.*, 1996), mRNA *in situ* hybridization reveals transcript distributions (Tabara *et al.*, 1996; Birchall *et al.*, 1995), and immunostaining demonstrates protein distributions. In addition, expression profiles, which lack the spatial resolution of expression patterns, have been generated by RT-PCR (reverse transcriptase polymerase chain reaction) based approaches (Grillo and Margolis, 1990; Johnstone, 2000), northern analyses (Alwine *et al.*, 1977) and, more recently, in genomic-scale microarray analyses (Kim *et al.*, 2001; Jiang *et al.*, 2001), all of which measure mRNA levels. In addition western analyses can provide expression profiles in terms of protein levels (Renart *et al.*, 1979).

3.2.1: Reporter Gene Fusion.

In the reporter gene fusion approach *C. elegans* is genetically transformed with a recombinant DNA containing a reporter gene fused to the promoter of a specific gene from *C. elegans* (Fire *et al.*, 1990). This fusion may involve ligation and cloning or can be achieved through PCR. When and where the reporter gene is expressed in *C. elegans* is directed by the promoter region to which the reporter is fused. The product of the reporter gene (e.g. β-Galactosidase or Green Fluorescent Protein) can be easily visualized and from inspection of transformed animals the location and temporal manner in which the *C. elegans* gene is expressed can be determined.

There are obvious caveats in the interpretation of reporter gene data. This has lead to caution in assuming that an expression pattern observed is an accurate reflection of the expression of the endogenous gene. For example, concern has been expressed that regulatory elements, crucial for precise expression of a gene, may not be contained within the DNA fragment that has been fused to the reporter. Frequently, in order to minimize these concerns, the DNA fragment used will be as large as possible or the reporter may be inserted into the gene being assayed so that DNA upstream and downstream from the point of fusion is also included. Neither of these precautions, however, guarantees that additional upstream, or even downstream, enhancers or silencers have not been omitted. There is

little information on the dispersal of regulatory elements for *C. elegans* genes in general as only a few genes have been subject to careful analysis in this regard (Thatcher *et al.*, 2001). Even though *C. elegans* protein coding regions tend to be more densely packed than in *Drosophila* (Adams *et al.*, 2000) or vertebrates they are much less densely packed than in yeast (Mewes *et al.*, 1997; Comeron, 2001; Poole *et al.*, 2003) and the frequency with which a *C. elegans* gene's regulatory elements may be found within or beyond adjacent genes is unknown. There are many examples of protein coding regions for distinct *C. elegans* genes being intertwined and it may be anticipated that intermingling by regulatory elements would be more readily tolerated. Failure of a reporter gene expression pattern to fully reflect expression of the endogenous gene might also arise from a gene having alternative promoters or splicing and the utilization of these alternatives gene structures could be affected by fusion to a reporter. Some *C. elegans* genes are regulated post-transcriptionally (Dalley *et al.*, 1993; Wightman *et al.*, 1993; Gaudet *et al.*, 1996; Puoti *et al.*, 2001; Puthalakath and Strasser, 2002) and this will not usually be apparent in reporter gene expression patterns. Study of this form of control can be performed using reporter genes (Sun *et al.*, 2003) but requires appropriate, informed design of the reporter gene fusion experiment.

The context of a reporter gene fusion, upon transformation of *C. elegans*, is different from that of the endogenous gene and this may also affect expression. The expression of reporter genes is inhibited in the germline (Kelly *et al.*, 1997; Kelly *et al.*, 1998). This was first noted from the repeated failure to see reporter gene expression for genes known, from other approaches, to be expressed in the germline. This inhibition of transgene expression may result from silencing of tandemly-repeated genes, the typical arrangement of exogenous DNA in the extrachromosomal arrays usually generated upon transformation of *C. elegans*. This inhibition appears to be relieved by increasing the complexity of the transforming DNA (Kelly *et al.*, 1997) or if the exogenous DNA is integrated into a chromosome in low copy number (Praitis *et al.*, 2001).

The juxtaposition of a reporter gene fusion to other DNA is also a consequence of *C. elegans* transformation and can result in inappropriate influences of irrelevant regulatory elements on the reporter gene fusion. Influences from a site of chromosomal integration

would be identified because normally multiple, independent *C. elegans* transformants are established and the site of integration is apparently random. Use of independent transformants also avoids problems from inappropriate juxtaposition of elements within the DNA being used in the transformation. Influences on the reporter gene fusion from the phenotypic marker gene used to identify the *C. elegans* transformants have been reported (Fukushige and Siddiqui, 1995), but the nature of these influences are known for the established markers and should not mislead.

Despite these reservations, the reporter gene fusion approach is widely used. The reporter gene fusion approach has many advantages and the vast majority of expression patterns in WormBase/ACeDB have been characterized using this approach. The technique is relatively easy to apply. Expression patterns generated have excellent spatial and temporal resolution and can be very firmly linked to an annotated gene in the genome. There may be a question over how well a reporter gene expression pattern relates to the function of the gene product, but at the very least it provides an accurate account of the transcription directed by the DNA fragment being assayed and the transcriptional control elements therein.

3.2.2: mRNA *in situ* hybridisation.

In situ hybridization provides gene expression patterns that are based on the distribution of specific mRNAs. A fluorescence *in situ* hybridization (FISH) procedure, with confocal imaging, was used in one study (Birchall *et al.*, 1995) to maximize signal over background. Sub-cellular resolution was achieved, allowing the distributions of mRNAs within cells to be visualized in intact animals at post-embryonic stages for several genes. For increased output, an alternative *in situ* hybridization procedure, with alkaline-phosphatase-based visualization in a 96-well, multi-well plate format, was developed (Tabara *et al.*, 1996). This approach has provided the thousands of expression pattern images presented in NEXTDB. The mRNA distributions for individual genes have also been determined using variations of these procedures.

3.2.3: Immunostaining.

Immunostaining, with an antibody raised against a gene product, can reveal a protein's distribution in the organism at all stages of development. Any restricted distribution of the protein within cells, to specific organelles or other sub-cellular structures, will also be revealed and may be relevant to interpretation of gene function. Confusion resulting from cross-reactivity of the antibody and observation of the distribution of non-target protein, can be avoided by comparing a wild type strain with a strain lacking a functional gene for the target protein. The main limitation of the immunostaining approach is the difficulty in producing antibodies and, as a consequence, relatively few gene expression patterns have been determined in this way.

3.2.4: Expression profiling.

Other techniques for observing gene expression in *C. elegans* do not provide the same resolution as the *in situ* techniques described above. For these profiling techniques, total protein or mRNA is prepared in bulk and then probed for the product of a specific gene (Reinke, 2000). Theoretically the protein or mRNA could be prepared from biological material of precise developmental age or anatomical structure. However, quite large quantities are needed and technical limitations means that for *C. elegans* the spatial and temporal resolution is poor. Synchronization of cultures allows the life-cycle to be divided little better than into six stages; embryos, the four larval stages and adults. Most of the key developmental processes are considered to occur during embryogenesis. The small size of the organism precludes physical dissection of tissue on the scale required and genetic dissection is utilized to generate populations of animals lacking particular tissues for comparison to populations of intact animals. Progress has been made recently in automated purification of specific cell types following dispersal of cells from embryonic stages and this promises to provide much better spatial resolution than had been possible previously (Zhang *et al.*, 2002).

DNA microarrays can be used to determine the mRNA levels for a large number of genes in parallel (Hill *et al.*, 2000; Reinke *et al.*, 2000; Jiang *et al.*, 2001; Kim *et al.*, 2001). DNA fragments for the majority of all identified *C. elegans* genes have been prepared and deposited, in order, onto glass slides to generate the microarrays. Fluorescently labeled

probes derived from mRNAs from two different samples can be simultaneously hybridized to a microarray, and the relative abundance of the transcript of every gene can be determined by comparing the signal intensities of each probe.

Before microarrays were developed, developmental mRNA profiles were generated for individual genes by RT-PCR (reverse transcriptase polymerase chain reaction) or northern analysis. In RT-PCR, total mRNA preparations were used as a template in quantitative PCRs with comparison to a constitutively expressed internal standard (Grillo and Margolis, 1990; Johnstone, 2000). In northern analyses, the total mRNA preparation was size fractionated by gel electrophoresis before detection of a particular transcript by hybridization with a gene-specific probe (Baugh *et al.*, 2001). In a similar way antibodies were used to detect levels of particular proteins in total protein preparations through western analyses (Renart *et al.*, 1979). All of these techniques have been used to generate data, that have appeared in the literature, of relevance to *C. elegans* gene expression.

3.3: Collation of published expression pattern data for ACeDB.

At the commencement of this research the Expression_Pattern class of ACeDB contained 245 gene expression patterns (several of which were duplicates). The majority of the expression patterns had been produced in the laboratory of Dr I. A. Hope at University of Leeds (UK). However many published reports containing gene expression patterns were not yet assimilated into ACeDB (version WS8, July 1999) (Stein and Thierry-Meig, 1999). As the research described later in this thesis was dependent on a complete listing of all *C. elegans* gene expression patterns it was necessary to perform the laborious process of identifying these patterns (Section 2.1). Therefore initially this chapter will describe a systematic literature survey in PubMed (http://www.ncbi.nlm.nih.gov/PubMed) and ACeDB to obtain all papers containing expression data, and the subsequent assimilation of these data into WormBase/ACeDB databases (http://www.wormbase.org). In addition, the research described in this chapter will also address concerns regarding the reliability of the expression pattern data produced by the reporter gene fusion approach. This will be achieved by comparing data produced by reporter gene fusion with the other experimental approaches and any inconsistencies will be evaluated.

3.4: Results.

3.4.1: Duplicated expression pattern data within ACeDB.

Initial examination of the expression pattern data in ACeDB (version WS8, July 1999) revealed nine duplications amongst the 245 entries at the time. These duplications were subsequently reported to the ACeDB/WormBase curators and resulted in the removal of the data from the database (Table 3.2).

Gene	*Duplicated gene expression patterns within ACeDB with their expression pattern identification numbers*
ZK899.4	Expr45 and Expr74
B0272.2	Expr40 and Expr69
B0272.4	Expr42 and Expr71
T14B1.2	Expr43 and Expr72
mec-3	Expr88 and Expr265
ggr-2	Expr244 and Expr245
aex-3	Expr203 and Expr335
glr-1	Expr247 and Expr249
cdh-3	Expr207 and Expr208

Table 3.2. **List of duplicated expression patterns in ACeDB.**

3.4.2: Searching for published *C. elegans* gene expression patterns.

After a rigorous survey of the literature, the number of gene expression patterns obtained totaled 202 (Appendix I). A small proportion (33 of 202) represented expression of the same gene but analysed by different experimental approaches (Table 3.3). The information on the expression pattern included a text description (including experimental details), the specific cell and/or cell groups, the sub-cellular localization, the life stage and level of expression. All extracted expression pattern data was communicated to Professor Paul Sternberg (responsible for WormBase, for inclusion into WormBase/ACeDB) and assimilated into WormBase in May 2001 (For example see information for expression pattern of the gene *tba-2,* [Figure 3.1]). The language used to describe each of the gene expression patterns was standardized and these text descriptions are fully searchable within ACeDB/WormBase thus allowing easy identification of related expression patterns (e.g. all genes that express in a specific cell or tissue). Each of the expression patterns within the database are linked to the corresponding genes within the sequence and physical genome maps.

Experimental technique	*Number of gene expression patterns*
Reporter gene fusion	83
Immunostaining	30
mRNA *in situ* hybridization	36
Northern blot analysis	39
Western blot analysis	3
RT-PCR	11

Table 3.3. The number of expression patterns obtained using the different experimental techniques.

Expression Pattern for tba-2

Summary

- Following hatching, 6 pharyngeal muscle cells stain becoming increasingly intense during larval and adult development. 6-7 neurons are stained in the ventral cord identified as DB motor neurons. Late larval and adult stages, staining of the set of DB and VB motor neurons in the ventral cord and their axonal processes along the ventral cord shows higher intensity of staining in late larval and adult stages. At L4, Pharyngeal muscles and the motor neurons show intense staining in L4 and adult stages. Fusion gene expressed in intestinal nuclei showing a gradient of expression, high staining in the posterior most intestinal nuclei in the tail region to a lower intensity of staining in the anterior intestinal nuclei. Intestinal staining decreases (L3-L4) and adult stages. Staining is detected in entire intestinal organ. Expression in neurons includes a set of DB and VB (19 cells) in the ventral nerve cord, a pair of posterior mechanosensory receptor neuron PLML and PLMR a single interneuron PVT in the pre-anal ganglion and a single neuron ALA in dorsal ganglion in the head.

- **Remark:** *tba-2 in this article is referred as alpha-2 tubulin, while the same authors referred it as tba-2 in later publications such as cgc2176. This information was extracted from published material (Archana Sharma-Oates, Andrew Mounsey and Ian A. Hope). Reporter gene fusion type not specified.*

More Details

Details

Expressed in:

Cell(s):	ALA I1L I1R I2L I2R I3 PLML PLMR PVT	view pedigree for these cell(s)
Cell Group(s):	neurons pharyngeal muscle ventral cord motor neurons	
Life Stage(s):	L1 larva L2 larva L3 larva L4 larva adult	

Expressed by:

Genetic Loci:	tba-2

Experimental Details

Reporter_gene Assay:	[tba-2::lacZ]. Used recombinant (alpha-2 tubulin 2.5 kb Pst1 genomic DNA fragment containing 2.1 kb of upstream sequence fused with lacZ) in the expression vector pPD16.51. --precise ends.

Bibliography

1993: Fukushige T et al. 1993. *Molecular cloning and developmental expression of the alpha-2 tubulin gene of Caenorhabditis elegans.*. Journal of Molecular Biology **234**: 1290.

Figure 3.1. The organisation within WormBase and ACeDB of the expression pattern data extracted from the literature for *tba-2*.

3.5: Comparison of expression pattern data.

3.5.1: Comparison of *C. elegans* gene expression patterns generated with reporter gene fusions, mRNA *in situ* hybridization and immunostaining.

Fifteen *C. elegans* gene expression patterns have been examined using the three techniques of immunostaining, mRNA *in situ* hybridization and reporter gene fusion (Genes are listed in Table 3.4 with details of expression patterns available from the WormBase database, [http://www.wormbase.org].) Although there are some differences between these descriptions (discussed below), the expression patterns generated by the three different techniques are largely consistent. The majority of differences can be attributed to fundamental characteristics of the approaches used to measure expression. Subcellular detail may be included in the gene expression pattern description generated by immunostaining because a precise distribution of the protein is revealed. For example all three techniques show expression of *csq-1* in differentiated body wall muscle cells but immunostaining demonstrates that its gene product, calsequestrin, is distributed in a "mesh-like" manner in these cells (Cho *et al.*, 2000). *asp-1* expression is in differentiated intestinal cells but the encoded protein, an aspartic protease, is restricted to the apical surface of these cells (Tcherepanova *et al.*, 2000). An interesting complication is presented with extracellular proteins. The two *C. elegans* type IV collagens, encoded by *emb-9* and *let-2*, (Graham *et al.*, 1997) and the type IV collagen binding protein, nidogen, encoded by *nid-1* were localized to body wall muscle basement membrane by antibody detection, and production localized to the body wall muscle cells according to the other techniques. Additional sites of expression of *nid-1* as demonstrated by the reporter gene fusion approach, may be *bona fide*, as in the animal ectopically-expressed nidogen has been shown to reach the correct extracellular sites and function appropriately (Kim and Wadsworth, 2000). For some gene products subcellular protein distribution can also be revealed with appropriately designed reporter gene fusion experiments. For example tagging of GFP to the C-terminal of the CCCH finger protein, PIE-1, demonstrated the same cellular and subcellular distribution in blastomeres of the early embryo as endogenous PIE-1, detected with PIE-1 specific antibodies (Reese *et al.*, 2000). Therefore the reporter gene fusion approach can provide detailed information regarding gene product distribution similar to immunostaining.

When the only inconsistency between three techniques is the apparent incompleteness of the pattern observed by *in situ* hybridization this probably reflects the relative insensitivity of this technique. For example the GATA factor encoding gene, *elt-2*, is expressed in the early E-lineage (Fukushige *et al.*, 1998). Reporter gene fusions and immunostaining reveal expression at the 2-E cell stage, while *in situ* hybridization does not demonstrate the presence of transcript until the 4 E cell stage. Presumably *in situ* hybridization is failing to detect the low levels of this transcript present when the gene is first transcribed. A similar explanation may apply for the *unc-64* expression patterns (Ogawa *et al.*, 1998; Saifee *et al.*, 1998). The syntaxins encoded by *unc-64* are primarily expressed in all neurons. Additional components in non-neuronal cell types, although weak and not detected by *in situ* hybridization, are presumably real as they are seen with both reporters and immunostaining.

A comparison of complex expression patterns generated using different techniques, as they are described in published studies can be particularly difficult. Some authors will claim unequivocally that a gene expression pattern "corresponds perfectly" to other published reports (e.g. *lin-26* in Dufourcq *et al.*, 1999). However, variation in the precise wording used by different authors to describe results can lead to considerable ambiguity. For example all descriptions for the myosin encoding gene, *unc-54*, include body wall muscle but there is variation in the details concerning which other muscle cells express this gene (Okkema *et al.*, 1993; Seydoux and Fire, 1994; Miller *et al.*, 1986). Alternatively, such differences in descriptions may simply reflect the different resolution that can be achieved with the different techniques. The gene *peb-1* encodes a DNA binding protein involved in pharyngeal development, but details of which pharyngeal cells the gene expresses in are only provided by reporter gene fusion and immunostaining experiments, presumably because resolution is poor with *in situ* hybridization (Thatcher *et al.*, 2001).

Finally discrepancy in maternal expression is a major inconsistency in the expression pattern descriptions for the 15 genes studied by immunostaining, mRNA *in situ* hybridization and reporter gene fusion approaches. Maternal expression pattern descriptions are confounded both by germline silencing of transgenes and by post-transcriptional control of gene expression. The *pal-1* transcript is maternally expressed and

can be detected by *in situ* hybridization in all blastomeres of early embryos (Hunter and Kenyon, 1996). Translational control, a key aspect of the function of this gene, restricts synthesis of the PAL-1 protein, a caudal homologue, to P2 and EMS, and their descendents. The gene is also expressed zygotically in the lineages of C and D, somatic blastomeres descended from P2, and this aspect of the *pal-1* expression pattern is revealed by reporter gene fusions (Edgar *et al.*, 2001). Maternal, germline expression is suppressed by silencing as discussed above (see Section 3.2.1). Similar considerations applied in examination of the early embryonic expression pattern of *glp-1* except that translation of the maternal transcript occurs in anterior rather than posterior blastomeres and immunostaining reveals the GLP-1 protein is localized to particular intercellular interfaces of the blastomere plasmamembranes (Evans *et al.*, 1994). Thus, although these gene expression pattern descriptions are not identical, they nevertheless accurately reflect different aspects of these genes' expression.

Similarly subtle differences in other expression pattern descriptions may also be physiologically appropriate. The *nid-1* gene expression is detected at the lima bean stage with immunostaining whereas it is detected earlier (at the late gastrulation stage) with mRNA *in situ* hybridization. This may be due to the time to translate the mRNA transcript into protein and then for the protein to reach a sufficient level for it to become detectable (Kim and Wadsworth, 2000; Kang and Kramer, 2000). The *in situ* hybridization result for *hlh-1* shows a lack of transcript between early and late embryonic components that is not seen with either reporter gene fusion or immunostaining approaches. HLH-1 protein may perdure through this temporal gap in the gene's transcription (Krause *et al.*, 1990; Seydoux and Fire, 1994).

For these 15 genes in the published accounts, consistency between the different expression pattern analyses would be expected. These descriptions result from studies focused on understanding the function of specific genes and experiments would have been pursued until anomalies had been explained and resolved. In addition inconsistencies resulting from experimental failings may have gone unpublished. However, the question of reliability remains for expression patterns determined less thoroughly, as in larger-scale gene expression pattern determination experiments (Lynch *et al.*, 1995; Birchall and

Albertson, 1995; Tabara *et al.*, 1996). An answer to this reliability issue may be found by restricting the comparison to just two of the three *in situ* approaches used for determining expression patterns for a particular gene.

3.5.2: Comparison of *C. elegans* gene expression data obtained by immunostaining and reporter gene fusion methods.

A comparison of gene expression data generated by both immunostaining and reporter gene fusion approaches, but not *in situ* hybridization, revealed that over a third (21 of 53) of the gene expression patterns were identical. Expression patterns were considered to be identical if expression is detected in the same cell groups, at the same life stages and at similar levels. Furthermore there are no gene expression patterns produced by these two approaches that appear to be distinctly different.

Three of the 53 patterns, those for *lmn-1*, *egl-32* and *hsp-16*, are identical apart from the previously acknowledged, germline expression problem (Section 3.2.1). Germline expression was observed using antibodies but not with the reporter gene fusions. There are four further genes, *hlh-2*, *nhr-2*, *unc-3* and *tlf-1*, where all aspects of the pattern are identical other than in the early embryo. In these patterns, expression as detected by the reporter gene fusion method is later in embryonic development than that detected by immunostaining. Again this could be due to the reporter gene fusion approach failing to detect expression in the germline. General zygotic transcription commences at approximately the 28-cell stage and earlier expression, as detected by immunostaining, is likely to be maternal (Newman-Smith and Rothman, 1998).

Almost all the remaining expression patterns (22 of 53), determined using just reporter gene fusions and immunocytochemistry, are consistent for some components, but distinct components are seen with at least one of the techniques. These additional components include a wide range of cell types including various muscle cells, the coelomocytes, various pharyngeal cells, neurons, hypodermal cells, the intestine or the spermatheca, and, as expected, the germline. Apart from the germline, the data is not sufficient to reveal any cell type as being particularly prone to being missed or inappropriately identified by either technique. A possibly less pronounced manifestation of this finding may be the differences

in the levels of expression, for different components of a gene's expression pattern, being inconsistent between these two approaches. The expression of the *ceh-6* gene in the excretory cell is observed at high levels with the reporter gene fusion approach but only weakly by immunostaining in comparison to other components (Bürglin and Ruvkun, 2001). Individual components of a complex expression pattern may be missed simply because of differences in the sensitivities of the techniques. The additional expression components for *tba-1*, detected only weakly with immunostaining, were not detected with a reporter gene fusion (Fukushige *et al.*, 1995). It is notable that for 19 genes the expression pattern generated by the reporter approach was missing a component seen using antibodies while the converse is true for only seven genes. Components are missing for both approaches for four genes.

Finally, for the remaining three genes in this category, *apr-1* (Hoier *et al.*, 2000), *lin-12* (Wilkinson and Greenwald, Herman *et al.*, 2000 and 1995, Levitan *et al.*, 2001) and *T03D8.1* (Zhang *et al.*, 2001), the depth of expression pattern descriptions for the two approaches are markedly different making comparison difficult.

3.5.3: Comparison of *C. elegans* gene expression data obtained by mRNA *in situ* hybridization and by reporter gene fusions.

Fewer gene expression patterns have been carefully characterized using *in situ* hybridization allowing fewer comparisons to be made with data produced using this technique. Nearly half (7 of 15) of the expression patterns analyzed by reporter gene fusions and *in situ* hybridization show full consistency between the two approaches. Another gene, *nhr-23*, gives identical expression patterns using the two approaches except for the absence of the germline component with the reporter gene fusion. For a further two genes, *lir-2* (Dufourqc *et al.*, 1999) and *R06C7.3* (Takemoto *et al.*, 2000), reporter expression is detected at a slightly later embryonic stage than with mRNA *in situ* hybridization. As discussed earlier (Section 3.2.1), this is probably due to maternal expression not being detected by the reporter gene fusion approach but could also reflect a delay between production of the mRNA and translation to detectable levels of protein product.

There are two genes, *F59B2.13* and *gap-2* (Hayashizaki *et al.*, 1998), which share common expression pattern components, however additional expression is detected in several cell groups by the reporter approach but not with the mRNA *in situ* hybridization approach. This is likely to be due to the insensitivity of the mRNA *in situ* hybridization approach. The same applies to three other genes, *fkh-1*, *nhr-25* and *spk-1*, but in addition these genes also give extra components by *in situ* hybridization that are not revealed by the reporter gene fusions. This implies that, like the *in situ* hybridization approach, the reporter gene fusion approach can miss expression pattern components but presumably, in this instance, because of failure to include transcription control elements in the constructed reporter gene fusions.

In summary, gene expression patterns produced by reporter gene fusion and *in* situ hybridization are largely consistent and there are no genes for which expression patterns appear distinctly different by these two approaches.

3.5.4: Comparison of *C. elegans* gene expression data obtained by immunostaining and by mRNA *in situ* hybridization.

Two thirds (8 of 12) of the expression patterns generated by both immunostaining and mRNA *in situ* hybridization are identical in terms of spatial and temporal expression. *In situ* hybridization detects the *nos-1* transcript before the 550-cell stage of embryonic development, the stage at which antibodies first detect NOS-1 protein (Subramaniam and Seydoux, 1999). The delay in detection of the NOS-1 protein, in comparison to the detection of the transcript, presumably simply reflects the time required to accumulate sufficient levels of protein for detection. The remaining three genes gave multi-component expression patterns, with an extra component seen using antibodies for *M01E5.5a/b* (Lee *et al.*, 2001), an extra component detected with *in situ* hybridization for *skn-1, (*Cox *et al.*, 1989; Maduro *et al.*, 2001*)* and an extra component seen with both techniques for *adm-1* (Podbilewicz, 1996). Based on the limited number of genes examined using both antibodies and *in situ* hybridization, these two techniques give highly consistent expression pattern data.

Gene	Approaches used to monitor C. elegans gene expression patterns		
	Immuno-staining	Reporters	In situ hybridization
asp-1			
ceh-13			
csq-1			
elt-2			
emb-9			
glp-1			
hlh-1			
let-2			
lin-26		Late	late
nid-1			
pal-1			
pie-1			
peb-1/T14F9.4		*****	
unc-54			
unc-64			
apr-1			
ceh-6			
ceh-17			
ceh-22			
dyn-1			
eat-20			
egl-5			
egl-8			
egl-32		*****	
elt-3			
F43D9.1			
flt-1			
ham-2			
hlh-2		late	
hlh-8			
hsp-16		*****	
ina-1			
itr-1			
K08F8.4			
kel-1			
let-23			
let-413			
lin-12			
lmn-1		*****	
ncs-1			
nex-1			
nhr-2		late	
odr-1/gcy-10			
odr-3			
osm-5			
osm-10			
pat-10			
rpm-1			
sad-1			
sel-1			
ser-1			
snb-1			
slo-2			
syd-2			
T03D8.1			
tba-1			
tlf-1			
tmy-1			
ubc-2			
unc-1			
unc-3		late	
unc-11			
unc-18			
unc-42			
unc-86			
unc-130			
vab-2			
vab-8			
B0464.5/ spk-1			
C47E12.8			
ceh-24			
elr-1			
F59B2.13			
fkh-1			
gap-2			
lir-1			
lir-2		late	
mab-5			
nhr-23		*****	

Table 3.4. **Genes for which expression patterns have been generated using two or more methods.**

Gene	Approaches used to monitor C. elegans gene expression patterns		
	Immuno-staining	Reporters	In situ hybridization
nhr-25			
R06C7.3		late	
unc-32			
vab-7			
adm-1			
apx-1			
gld-1			
let-805			
M01E5.5a/b			
mex-1			
mex-3			
nos-1	Late		
nos-2			
pos-1			
skn-1			
T02G5.8			

Table 3.4 continued. **Genes for which expression patterns have been generated using two or more methods.**

Key	
	Identical pattern
	Level of expression different
	Missing components
	Additional components
	Additional and missing components
	Slightly different
	Comparison not possible
	Indistinguishable pattern between the two techniques
*****	Germline components missing
Late	Late onset of expression

3.6: Comparison of temporal gene expression data.

The techniques of RT-PCR, northern and western analyses allow the developmental timing of a gene's expression pattern to be determined but, for *C. elegans*, cannot provide spatial information. Nevertheless comparison could be made to the temporal aspects of the developmental expression pattern descriptions generated using the three main, microscope-based techniques, although frequently, in these descriptions, temporal aspects have been given lower priority. *C. elegans* genes examined by RT-PCR, northern or western analysis are listed in Table 3.5.

3.6.1: Comparisons with RT-PCR analyses.

Approximately half (8 of 15) of the genes examined with both reporter gene fusions and RT-PCR have the same temporal gene expression profiles. For two of the remaining seven

genes, *cyp-4* and *gar-1*, there are no temporal aspects provided in the expression pattern descriptions generated with the reporter gene fusion approach and therefore no comparisons can be made. The differences in temporal expression profiles of the other five genes are quite minor.

For *med-1*, expression was first detected by RT-PCR at the 4-cell stage, when EMS is first formed (Maduro *et al.*, 2001). Using a reporter gene fusion, *med-1* expression was detected in EMS but not until the 6-cell stage. This probably reflects the delay between transcription and translation. For *col-12* and *lin-42* expression is observed at the same life stages by both methods but there are differences in the accounts of changes to the level of expression (Johnstone and Barry, 1996; Joen *et al.*, 1999). Only the reporter gene fusion approach revealed the level of *lin-42* expression to be higher in animals undergoing ecdysis. For *col-12*, an increase in the level of expression at the L4/adult molt was detected by RT-PCR but not by the reporter gene fusion approach. Expression of *gly-14* is detected postembryonically by both procedures, but is only detected by RT-PCR at the embryonic stage (Chen *et al.*, 1999). This may be due to germline or weak expression in the embryo that is not detected by the reporter gene fusion. The expression profile of *dpy-7* extends into the adult for the reporter gene fusion method, but not for RT-PCR, presumably because the protein persists longer than the transcript (Johnstone and Barry, 1996).

There are only two genes, *wee-1*(Wilson *et al.*, 1999) and *alt-1* (Gomez-Escobar *et al.*, 2002), for which expression patterns have been generated by both mRNA *in situ* hybridization and RT-PCR methods. For *wee-1* temporal expression is the same for both approaches. For *alt-1* the temporal expression profiles differ in that expression is observed only in embryos and adults by *in situ* hybridization but is also in the larvae according to RT-PCR. It is possible that the level of expression in larvae is weak and is therefore detectable by RT-PCR but not by *in situ* hybridization.

There are also only two genes, *egl-32* and *osm-5*, for which expression patterns had been generated by immunostaining and RT-PCR however as neither of the protein distribution descriptions contained temporal information it was not possible to make comparisons.

3.6.2: Comparisons with northern analyses.

Almost all (7 of 8) of the temporal patterns generated by the RT-PCR and northern blotting approaches are identical. The transcript for *Y60A3A.12/chk-2* (Oishi *et al.*, 2001) was detected, by northern analysis, at all larval stages except L2, when RT-PCR was able to detect the transcript, albeit at low levels. Again this is most likely due to RT-PCR being more sensitive.

The majority (7 of 9) of the expression profiles generated with both immunostaining and northern analysis could not be compared due to either very little or no temporal information provided for the immunostaining descriptions. One gene, *nos-3* (Kraemer *et al.*, 1999), had an identical temporal profile as demonstrated by the two approaches. For the other gene for which expression profiles can be compared, MES-1 protein appears to be restricted to the embryos by immunostaining, while the *mes-1* transcript is detected in late larval and young adults by northern analysis (Berkowitz and Strome, 2000). For *mes-1*, the transcript is deposited in the germline maternally and is not translated until after fertilization.

There are only eight expression profiles generated with mRNA *in situ* hybridization and northern analysis. Two of these, for *nhr-25* and *csq-1*, are identical and the others could not be compared because of insufficient temporal information provided in the descriptions of the in *situ* hybridization patterns.

Just under half (9 of 22) of the gene expression profiles generated using northern analysis and reporter gene fusions appear to be identical. Nine of the remaining 13 could not be compared because of the lack of temporal information provided in the descriptions from the reporter gene fusion approach. However, there are two genes, *daf-7* and *egl-27*, where expression is not observed at a specific life stage by northern analysis, but is detected by the reporter approach. For *vab-7*, the reverse is observed (Ahringer, 1996). The latter may be explained by the problems of reporter gene fusions with germline expression, while the former may reflect the greater sensitivity of the reporter approach. In addition, there is one gene, *uvt-4*, where the temporal expression profiles match for the two methods, but the changes in levels of expression do not directly correspond.

3.6.3: Comparisons with western analyses.

There are 13 *C. elegans* genes for which an expression profile has been determined by western analysis. However, lack of temporal information prevents comparison for 5 genes; *fox-1*, *hsp-25*, *ncc-1*, *par-1* and *T03D8.1*. The western expression profiles are identical to the profiles generated by reporter gene fusion analysis for *daf-1*, *fem-1*, *lap-1* and *tba-1*; by northern analysis for *bir-1*, *fem-1*, *tba-1* and *T02G5.8*; by immunostaining for *tba-1*, and by RT-PCR for *lap-1*. There are just two genes for which the western expression profile disagrees slightly with expression profile data from other approaches. For *ins-18* protein is not detected by western analysis at the L1-stage and in adults, however expression is detected at these stages by reporter gene fusion analysis, although the reasons for this distinction are not apparent (Gregoire *et al.*, 1998; Kawano *et al.*, 2000). For *F46H6.1* the difference is only in the level of protein versus mRNA (Kawano *et al.*, 2000). By the western approach the level of the protein varies between the different developmental stages whereas with the northern approach the level of expression remains constant.

Genes	**Approaches used to monitor *C. elegans* gene expression patterns**					
	Immnu-staining	Reporters	*In situ* hybridization	RT-PCR	Northern	Western
alt-1						
chk-2						
col-12						
cyp-3						
cyp-4						
cyp-8						
daf-12						
dpy-7						
elt-3						
egl-32						
gar-1						
gar-2						
gly-12						
gly-14						
lap-1						
lin-42						
med-1		late				
osm-5						
sqt-1						
wee-1						
csq-1						
daf-7						
dbl-1						
egl-27						
F55A8.2a						
fkh-1						
ftt-1						
ftt-2						
glp-1						
gly-13						
hbl-1						
let-502						
lin-12						
mef-2						
mel-11						
mes-1						
mog-1						
myo-3						
nhr-2						
nhr-25						
nos-3						
tba-2						
ubq-2						
unc-15						
uvt-4						
vab-7						
bir-1						
daf-1						
fem-1						
fox-1						
hsp-25						
ins-18						
ncc-1						
par-4						
rhi-1						
T02G5.8						
T03D8.1						
tba-1						

Table 3.5. Genes for which expression profiles have been generated using two or more methods.

Key	
	Identical pattern
	Level of expression different
	Slightly different
	Comparison not possible
*****	Germline components missing
Late	Late onset of expression

3.7: Discussion.

This chapter describes the systematic identification and collation of all published *C. elegans* gene expression pattern data prior to July 2001. This process was performed for several reasons. Firstly, a more complete listing of all published gene expression patterns was essential for the research described later in this thesis (Chapter 4). Secondly, to address the occasional questioning of the reliability of the reporter gene fusion approach for determining gene expression patterns. Thirdly to become familiar with *C. elegans* developmental anatomy and gene expression pattern descriptions. Finally, to draw together expression pattern data into single database ACeDB/WormBase.

Two hundred and two gene expression patterns were identified that had not previously been assimilated into WormBase/ACeDB. Analysis revealed that there were no major differences in the data produced by the various methods. Although approximately 30% of gene expression patterns were identical (Section 3.4), there were some differences between the results of the different approaches which warrant discussion:

1) In general, the resolution of images obtained using the reporter method was of higher quality than those using *in situ* hybridization (from NEXTDB) and this is reflected in the more detailed descriptions. Furthermore the published descriptions of patterns generated by reporter gene fusion contain a greater degree of detail than those determined by *in situ* hybridization possibly suggesting that the latter technique is less sensitive.

2) Maternal gene expression (germline) is generally not observed with the reporter gene approach, (as is consistent with many previous observations; Kelly *et al.*, 1997, Kelly and Fire, 1998 and Seydoux and Strome, 1999). However, for unknown reasons, expression of *egl-32* is detected in the germline using reporter gene fusion (GFP) (Miguel-Aliaga *et al.*, 1999).

3) There are a number of gene expression patterns in the public domain that could not be used in this study. This is due to the following reasons:
i) Several published papers describe multiple gene expression patterns generated using two or more methods but fail to state which technique was employed for which pattern.

ii) In some published papers expression patterns generated with one technique (usually immunostaining) are described in much greater detail than expression patterns generated with an alternative approach. As a consequence comparisons between the two are difficult.
iii) Many of the reporter, mRNA *in situ* hybridization or the immunostaining based expression patterns contain little or no temporal information. Consequently it is not possible to compare these data with those generated by RT-PCR, northern and western analyses.

4) Since this analysis was performed large quantities of data on transcript levels have been accumulated through microarray analysis. A comparison of these data with expression patterns generated with other approaches would be worthwhile but is a substantial project in itself. In addition there is now a substantial body of mRNA *in situ* hybridization data available from NEXTDB. However, the lack of annotation of these data and the difficulty of interpretting captured digital images would make comparisons difficult.

In summary a comprehensive analysis of *C. elegans* gene expression data has been performed to evaluate the consistency of patterns generated by different experimental techniques. Generally there is substantial agreement between the different approaches used and most of the inconsistencies can be attributed to well-recognised limitations of a particular approach, however immunostaining and reporter gene fusion provide the best resolution of *C. elegans* gene expression. Currently there are 2045 published gene expression patterns assimilated in WormBase and this has been possible due to a well-funded and coordinated curation project.

Chapter 4

Chapter 4: The identification of *cis*-acting elements from the promoter regions of co-regulated *C. elegans* genes using MEME software.

4.1: Introduction.

As discussed in the General introduction (Section 1.7) the detection of *cis*-acting regulatory regions is potentially an important approach with which to understand genome regulation. However the identification of putative regulatory motifs is fraught with difficulties due to the unknown size, variability of sequence and ill-defined location, inherent to these elements. A number of different algorithms have been developed for the identification of *cis*-acting elements. In this thesis two fundamentally different approaches will be employed. In this chapter, the so-called "alignment" method (which MEME uses) will be employed, whereas in Chapter 5 results obtained using the "enumerative" or "exhaustive" method (which SPEXS uses) will be described (Ohler and Niemann, 2001).

The software MEME was initially selected because it was the most widely-used, fully automated software for motif detection in bacterial and yeast studies, although, not previously used for more complex organisms such as *C. elegans*. The "alignment" approach, upon which MEME is based, aims to identify motifs by local multiple alignment of all sequences (Ohler and Niemann, 2001). Direct multiple alignment is a computationally demanding process and a number of different algorithms, forerunners of MEME, were developed to perform this task. One such early algorithm was EM (see Section 1.8), a pseudo-code description of the basic EM algorithm is given below:

```
1.    EM (dataset, W) {
2.        choose starting point (ρ)
3.        do {
4.            reestimate z from ρ
5.            reestimate ρ from z
6.            } until (change in ρ < ε)
7.            return
8.    }
```

The motif is defined by:

$\rho = \{ \rho_j^k; j=1, \ldots, W; k=1,\ldots 4\}$

Where ρ_j^k is the probability that base k occurs at position j of the motif. Its occurrence in the sequence is defined by $z = \{z_i | i=1, \ldots, N-W\}$,
where Z_i is the probability the motif defined by ρ starts at position i in the sequence, and N the length of the sequence. The algorithm begins with an estimate of ρ (which might be generated randomly, or from the sequence start).

EM starts with a random estimate of the motif model description, ρ, and then estimates the probability of each possible starting point of the example motif in the sequences in the dataset (*z*). However there were a number of disadvantages inherent to this algorithm;

i) Initial data sets selected with little consideration may result in the EM output converging inappropriately (to a local minimum).

ii) It assumes that each sequence contains exactly one occurrence of the motif. (If all sequences do not contain the motif it is not possible for EM to identify a consensus motif.)

iii) The output only contains the most conserved motif and no other less conserved motifs.

The MEME algorithm improves upon the capability of EM and overcomes the deficiencies described above (Bailey and Elkan, 1995a). MEME is a tool that can perform unsupervised learning i.e. it takes as input a set of sequences and identifies a pattern that is common to some of the sequences (Bailey and Elkan, 1995a). MEME overcomes the limitations of EM by systematically selecting motif starting-points, based on a training data set. It allows searching with the EM model (one occurrence per sequence [OOPS]) and also gives the option to select a different model which allows for zero (zero or one occurrence per sequence [ZOOPS]) or multiple occurrences of the motifs in a sequence (two-component mixture [TCM]). In addition MEME erases the appearance of a motif after it is found and continues searching for additional common motifs (Bailey and Elkan, 1995a). Potential disadvantages of MEME are that the number of different motifs to be identified in the data set and the width of the motifs has to be specified by the user. However, a major advantage of MEME is that the TCM model is less influenced by noise, and in addition, motifs can be detected if as few as 20% of the sequences contain the motif (Bailey and Elkan, 1994). MEME has previously been applied to the identification of protein and DNA motifs in bacteria and yeast (Mount, 2001; Bailey and Elkan, 1995b; Bailey *et al.*, 1997), and it was therefore considered a potentially ideal software for detecting motifs from a set of *C. elegans* genes that are considered to be co-regulated.

4.2: Results.

4.2.1: MEME analysis using the upstream regions of genes expressed in the excretory cell.

MEME was first executed using a file containing the entire region located upstream from the translational start site of all 16 genes expressed in the excretory cell. The excretory cell was selected for this analysis for reasons stated in Section 1.10. The proteins encoded by the genes expressing in the excretory cell used in this and other subsequent analyses fall into three distinct categories; 1) approximately 40% of the genes encode for enzymes involved in signalling cascade, 2) with a fifth of the genes products implicated in ion transport and 3) the next highest number of genes encode for calcium binding proteins and proteins that may play a role in transcription. Furthermore approximately 14% of the genes have not had their protein products classified thus far (Figure 4.1). Typically *cis*-acting elements occur within the non-coding regions of DNA (Hughes *et al.*, 2000) and more specifically up to 1-2 kb upstream from the translational start (Zhang *et al.*, 2002; Frith *et al.*, 2001) although some elements may occur outside this region. To minimize the chances of missing the elements that are beyond the 2 kb region, the entire upstream region to adjacent gene was analysed. Figure 4.2 demonstrates a typical example of MEME output generated for each motif identified by the software.

F33D4.2a/itr-1	Ion transport protein
B0285.6	Sodium/sulfate symporter
F41E7.1	Sodium/hydrogen exchanger
R05G6.6	Sodium:dicarboxylate symporter
R12G8.2	K+ channel, pore region
T06F4.2/clh-4	CBS domain, Cl- channel, voltage gated, Transmembrane amino acid transporter protein, Amino acid/polyamine transporter, family IIAromatic amino acid permease
Y18D10A.23	Transmembrane amino acid transporter protein, Amino acid/polyamine transporter, family IIAromatic amino acid permease
Y70G10A.3	Organic anion transporter polypeptide (OATP), C-terminal
ZK455.7/pgp-3	ABC transporter, transmembrane region
C17H12.14	H+-transporting two-sector ATPase, E subunit
F35H10.4/vha-5	V-type ATPase
Y38F2AL.3/vha-11	V-ATPase subunit C
Y46G5A.4	AAA ATPaseDEAD/DEAH box helicaseHelicase, C-terminal
R10E11.2/vha-2	Vacuolar H+-transporting two-sector ATPase, C subunit
R10E11.8/vha-1	ATP synthase subunit C, H+-transporting two-sector ATPase, C subunit,Vacuolar H+-transporting two-sector ATPase, C subunit
T14E8.1a	Tyrosine protein kinase
T17E9.1a/kin-18	Eukaryotic protein kinaseSerine/Threonine protein kinase, Tyrosine protein kinase
C09B8.7/pak-1	Eukaryotic protein kinaseSerine/Threonine protein kinase, Tyrosine protein kinase
C46C2.1	Eukaryotic protein kinaseSerine/Threonine protein kinase, Tyrosine protein kinase
Y75B12B.5/cyp-3	Peptidyl-prolyl *cis*-trans isomerase, cyclophilin type
C33D9.1/exc-5	RhoGEF domain, PH (pleckstrin homology) domain, Pleckstrin-likeDH domain, Zn-finger, FYVE type
F10B5.1	Ribosomal protein L10E
F54C9.1	Eukaryotic initiation factor 5A hypusine, DNA-binding OB fold
F44A2.5	Initiation factor eIF-4 gamma, middle
K02B12.1/ceh-6	POU homeobox
R13H4.5	Cyclin-like F-box
F55C7.7/unc-73	Src homology domain 3, RhoGEF domain, Fibronectin type III domain, PH (pleckstrin homology) domain, Cellular retinaldehyde-binding (CRAL)/Triple function domain (TRIO), Guanine-nucleotide dissociation stimulator, CDC24Basic helix-loop-helix dimerization domain, bHLHImmunoglobulin subtypeImmunoglobulin-likePleckstrin-likeDH domain, SH3 domain, Spectrin repeatVitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
K03D10.1/kal-1	Whey acidic protein, core region
T08G11.2/egl-32	SH2 motif
F04F6.1/clh-3	Nematode 7TM chemoreceptor (probably olfactory)
R107.8/lin-12	EGF-like calcium-bindingLaminin-type EGF-like domainNotch (DSL) domain
R06A10.2/gsa-1	Guanine nucleotide binding protein (G-protein), alpha subunit
R31.1/sma-1	Actinin-type actin-binding domain containing proteins
F45E10.1/unc-53	Actinin-type actin-binding domain containing proteins Calponin-like actin-binding
T21B10.5	Nuclear protein SET
Y62E10A.1	60S Acidic ribosomal proteinRibosomal protein P2
C14A4.12	unclassified
F54C9.7	unclassified
K04G2.8a/apr-1	unclassified

Figure 4.1. The protein products encoded by the excretory cell-expressing genes.

```
********************************************************************
MOTIF  1            width =  12         (12 bp) sites = 7.0  (expected number of occurrence in the given
sequences)
********************************************************************
Simplified      A  ::::2:1:::::
motif letter-   C  39:984655953
probability     G  7::::43:4::6
matrix          T  ::9::2:4::4:
Above is the observed frequency of each nucleotide in each column of the matrix.  The columns represent each nucleotide within the motif.  The frequency values
are multiplied by 10 and rounded-up to the nearest integer.  Zeros are represented as colons.

          bits 2.2
               2.0
               1.7
               1.5
Information    1.3  * *      *
content        1.1  ***      *
(11.7 bits)    0.9  ****     *
               0.7 *****  *****
               0.4 ************
               0.2 ************
               0.0 ------------

Multilevel          GCTCCGCCCCCG
consensus           C    CGTG TC
sequence
This figure represents the information content of the motif.  It illustrates the degree of degeneracy at each position within the motif.  The lower the information
content value the greater the degeneracy at the particular position.  The scale is logarithmic to the base 2 (bits).  The sum of all the columns is shown on the Y-axis.
The multilevel consensus sequence shows all bases in each column of the motif that occur with a frequency of >0.2.
```

Figure 4.2. An example of MEME-generated output for one predicted motif. Red text provides explanations for each of the MEME outputs.

```
--------------------------------------------------------------------------------
        Motif 1 in BLOCKS format
--------------------------------------------------------------------------------
BL   MOTIF 1 width=12 seqs=7
y62e10a.1                ( 1813) GCTCCGGTCCCG  0.991293
f44a2.5                  ( 4567) GCTCCCACGCTG  0.737914
y70g10a.3                ( 1033) CCTCCCCCCCCC  0.995644
c14a4.12                 ( 3715) GCTCCGCCCCCG  0.998438
f10b5.1                  ( 1300) GCTCCTGCGCTC  0.983271
f54c9.1                  ( 1133) CCTCAGCTCCTG  0.946726
r05g6.6                  ( 2121) GCTCCCCTGCCG  0.999503
//
```

The predicted locations of the motif in each sequence and the probability that the motif starts at that location. The sites reported depend on the motif model used: (1) OOPS, the most probable location in each sequence is given; (2) ZOOPS, the most probable location in each sequence is reported but only probabilities greater than 0.5 (a significant level for Baysian statistics); (3) TCM, all positions in each sequence with probabilities >0.5 are shown.

```
--------------------------------------------------------------------------
        Possible examples of motif 1 in the training set
--------------------------------------------------------------------------
Sequence name             Start  Score              Site
-------------             -----  -----            ------------
y62e10a.1                  1813  16.86 GCAATTGTGG GCTCCGGTCCCG CAGGTAATAA
f44a2.5                    4567  14.84 GAGCAGTTTT GCTCCCACGCTG TATTGGTCTA
y70g10a.3                  1033  16.20 TTAACAATGA CCTCCCCCCCCC ATAATGTTAT
c14a4.12                   3715  18.39 CGGAAAAAAT GCTCCGCCCCCG AACCATGGGT
f10b5.1                    1300  13.66 AGGATACACA GCTCCTGCGCTC GTCCCAAGGC
f54c9.1                    1133  13.33 GTTTTTGTTG CCTCAGCTCCTG CTTAAGTTTT
r05g6.6                    2121  17.45 CCTCCATCAT GCTCCCCTGCCG AACCAGCTTC
--------------------------------------------------------------------------
```

Above are all the possible examples of motif 1 in the data set, obtained using the TCM option. The list is based on using a position-dependent scoring matrix (log-odds matrix) to search each sequence. The threshold score for displaying a site is chosen such that the expected number of incorrect assignments will equal the expected number of missed but correct assignments. Positions before and after the motif are also shown.

```
log-odds matrix: alength= 4 w= 12 n= 41864 bayes= 12.5458
 -2.981   0.310   1.514  -2.915
 -2.990   2.015  -2.987  -2.815
 -2.860  -2.275  -2.939   1.728
 -2.855   2.013  -2.923  -2.942
 -0.856   1.805  -2.795  -2.918
 -2.969   0.845   0.828  -0.749
 -1.120   1.317   0.293  -2.854
 -2.799   1.182  -2.986   0.675
 -2.986   1.254   0.809  -2.979
 -2.986   2.014  -2.908  -2.838
 -2.992   1.266  -2.944   0.585
 -2.976   0.451   1.452  -2.955
```

This is the position specific scoring matrix. This matrix is a log-odds matrix calculated by taking the log (base 2) of the ratio of the observed to expected counts for each nucleotide in each column of the profile. The four columns represent the bases ATGC and each row represent a different position in the motif. The counts for each column may have additional pseudocounts added to compensate for zero occurrences of a nucleotide in a column or for a small number of sequences.

```
letter-probability matrix: alength= 4 w= 12 n= 41864
 0.035679  0.275167  0.653707  0.035447
 0.035476  0.897668  0.028875  0.037981
 0.038821  0.045884  0.029837  0.885458
 0.038945  0.896086  0.030183  0.034786
 0.155674  0.775985  0.032981  0.035360
 0.035982  0.398700  0.406304  0.159014
 0.129657  0.552968  0.280392  0.036982
 0.040497  0.503758  0.028896  0.426850
 0.035577  0.529646  0.400863  0.033914
 0.035575  0.896544  0.030487  0.037394
 0.035408  0.533929  0.029749  0.400914
 0.035808  0.303513  0.626211  0.034468
```

The motif base-frequency provided demonstrates the frequency of the nucleotide found in each column profile. The four columns represent the bases ATGC and the rows represent a different position in the motif. The notations "alength" represents the alphabet used (i.e., DNA), "w-12" is the width of the motif, and "n" is the number of characters in the sequences.

Figure 4.2 continued. An example of MEME-generated output for one of the predicted motifs. Red text provides explanations for each of the MEME outputs.

For the initial analysis the outputs when using the following options (Section 2.3), were considered:

i) –mod options (motif distribution): OOPS, ZOOPS and TCM.

ii) –W (motif widths in bp): 8, 12, 15 and 20.

iii) –nmotifs (number of motifs required for output): 10, 20 and 50.

First, MEME was executed using the options –W 12, -nmotifs 10 and each of the three different motif distribution options OOPS, ZOOPS and TCM. The detected motifs were ranked by MEME in order of conservation within the sequences from one (the highest) to 10 (the lowest). Although most of the motifs generated by the OOPS model were simple and AT-rich (Table 4.1, OOPS column), motifs (1 and 10) were more complex with a higher GC content. However, these two motifs show considerable degeneracy with 7/12 ambiguous positions for motif 10 and 8/12 for motif 1. Motif 10 is in fact part of the repeat element, CeRep10 and therefore was unlikely to be a transcriptional control element.

Out of the ten motifs detected by the OOPS and ZOOPS options, six motifs were present in both the sets. Motifs identified using the ZOOPS and OOPS option were either simple and AT-rich or showed a high level of degeneracy.

Using the TCM option a completely different set of motifs was identified. Nevertheless again these motifs were mostly AT-rich and with low information content (Tables 4.1 and 4.2). It is likely that OOPS and ZOOPS identified a similar set of motifs because both options detect one motif per sequence whereas in contrast TCM allowed for any number of occurrences within a sequence. As a consequence different motifs were identified when the TCM option was specified.

Therefore for all subsequent analyses it was decided that the OOPS option should no longer be used. OOPS generates similar motifs to ZOOPS and does not allow for the absence of the motif within a gene, a distinct possibility. The TCM model was considered the best option given that it allows for multiple occurrences of the motif which is the typical arrangement for eukaryotic transcriptional control elements (Spieth *et al.*, 1985).

OOPS	*ZOOPS*	*TCM*
GCTCCCCCCCTG C TGATG CC	GGTCTCGTCCCG CCG A CA A	AATTTTCAAAAA TTA GGGG T T
TTTTCCAATTTT G C G	TTTTCCAATTTT G C	TTTTTCATTTTT CCTTA C AG C
AAAATCTGAAAA T G	GGCGAAAACGGG TG GG CTC A T	AAAAAAGAAAAA G G TAGGG
CTGCTGAAAAAT T C G T G	AAAATCTGAAAA GT	GCCCGCGTGCCG CATGTTTCCTTC G CGCG
AAAATTGAAAAA A CG	GAAAAGTCGGAA A A A C G	TTTTAATTTTTT CCTTA C A
TTCCAATTTTTC A T G	CTCCCTCTCCCA GC C TG G	TTTTTCAATTTT CAGTA G
TTTTCAAAAAAA G G	TTTTTTTTCATT AG	AAAAATTAAAAA TAAGG CT
TTTTTTTTCATT C G	TTTTTGTTGAAA T G C	***CCGATTTGCCGG*** ***G AG T AA***
AATTGCAAAAAA C G C	TTTTCAAAAAAA CG T	***GCCGATTTGCCG*** ***G AG G***
GGCGAAAACGGG G GG CTC A A T	TTCCGATTTTTC CG AG GG A C	***GCCGATTTGCCG*** ***G AG T***

Table 4.1. The 10 motifs (in descending order) detected in the upstream region of the 16 excretory cell-expressing genes. The data were generated using the MEME options –nmotifs 10, -W 12 and –mod OOPS, ZOOPS or TCM. Similar motifs detected by the different options are highlighted in the same colour. The last three motifs in the TCM column (emboldened and italicised) were largely identical.

The majority of motifs generated from the above analyses were of low sequence-complexity and AT-rich. This is not unexpected since the *C. elegans* genome is known to be AT-rich particularly in non protein coding regions (Pilgrim, 1998; Surzycki and Belknap, 2000). Whether these motifs are true *cis*-acting elements, given that they are simple runs of As and Ts, is questionable. However previous analyses have shown that *cis*-acting regulatory elements can be of low sequence complexity, for example the hox element (TAATNN, [Grant *et al.*, 2000]) and therefore the motifs generated required

further investigation. This investigation demonstrated that most of the AT-rich motifs were part of tandem repeats and therefore unlikely to be *cis-acting* elements. However, motif number 1 from the OOPS group, motifs 3, 5 and 6 from the ZOOPS group and motifs 4, 8, 9, 10 from the TCM group (Table 4.1) were not AT-rich. When these motifs were used to search in ACeDB (local version), the fourth and the last three motifs from the TCM set were part of the repeat element CeRep3. In addition, the fourth motif from the TCM set was also part of a "dispersed repeat" CeRep22. Motif number 4 occurs in the intergenic region of 6 of the 16 gene sequences (Figure 4.2), with multiple occurrences in two of the sequences *Y62E10A.1* (eight occurrences, Figure 4.3) and *F44A2.5* (three occurrences). Similarly motif 10 occurs with a frequency of 30 in the gene *Y62E10A.1*. Motifs 8, 9 and 10 appear to be almost identical (Table 4.1). The last motif from the OOPS set (also detected in ZOOPS) was part of the repeat element CeRep10. Motif 1 from the OOPS set was contained in very few sequences as was the case with motifs 5 and 6 from the ZOOPS set.

Possible examples of motif 4 in the training set

Sequence name	Start	Score	Site
f44a2.5	1752	8.70	TGGCACGGAC GCTCGGTCGCCG TTGCCAGCAA
f44a2.5	4498	9.00	CCAATTCACC GCCCCCCTGCTC TTATCCGTTC
f44a2.5	5499	8.62	TGTCCTGATC GCCCTCTTGCAC TTTTCGAACC
c14a4.12	7952	8.64	TTCGCGACGA GACCCGGTGCCG TATTTCTGGG
f41e7.1	1361	9.41	AAACACGCCG GCTCTCCTCCCG TTTATTCCGT
r05g6.6	2121	9.62	CCTCCATCAT GCTCCCCTGCCG AACCAGCTTC
y18d10a.23	2329	10.02	TTTGAGATTT GCCGGTTTGCCG GAAATTTTCA
y62e10a.1	244	8.75	TACTGTCGAG GCGCGCGTGGCG AGACCCACTC
y62e10a.1	1813	8.93	GCAATTGTGG GCTCCGGTCCCG CAGGTAATAA
y62e10a.1	5995	10.02	CCTGCTATTT **GCCGGTTTGCCG** ATTTGCCGAA
y62e10a.1	6029	10.02	TTTGCAATTT **GCCGGTTTGCCG** GAATTTGACA
y62e10a.1	6134	10.02	CCGGCAATTC **GCCGGTTTGCCG** GAATCGGTCA
y62e10a.1	6266	9.06	GTTCAATTAA **GACCGTTTGCCG** GTTTTCCGAT
y62e10a.1	6379	10.02	CCGGCAATTT **GCCGGTTTGCCG** ATTTCCGAAA
y62e10a.1	6521	8.58	TCGTCGATTT **GCCGGTTTGTCG** GAAATTTAAA

Figure 4.2. Genes containing multiple copies of the fourth TCM motif GCCCGCGTGCCG (see Table 4.1) identified by MEME. This motif is part of the repeat element CeRep 3, located in multiple copies upstream from the start of gene *Y62E10A.1* (Figure 4.3). The sites emboldened were also detected as TCM motifs 8, 9 and 10 (CCGATTTGCCGG) (Table 4.1).

Executing MEME with the options –mod TCM, -nmotifs 10 and motif widths of either 8, 12, 15 or 20 bp generated similar sets of motifs (Table 4.2). The majority of the

motifs generated using the width 12 were also contained within the larger motifs (widths of 15 and 20). Using the –W 8 option (width of 8 bp) the last three motifs were completely non-specific and out of the seven specific motifs, the motifs ranked third, fifth and seventh were almost identical. The highest-ranking motif from the 8 bp set was also in the 12 bp set and the sixth motif was in the 15 bp set at the same ranking (Table 4.2). In the 8 bp set, only one motif was GC-rich however all but one of the positions within the motif had a high degree of degeneracy. In addition, the fourth motif from the 12 and 15 bp sets and the fifth motif from the 20 bp set, did not contain any conserved positions within the motif. Furthermore, the results generated from the –W 20 option contained six (out of ten) motifs that were simple runs of As and Ts and therefore very unlikely to be candidate recognition elements. Moreover, the fourth motif GCCCGCGTGCCG from the 12 bp set was found in the repeat element CeRep22 (Table 4.2). Similarly the last three motifs from the 12 bp set had a high GC content and were also detected within the 15 and 20 bp sets but were found to be part of the repeat element CeRep3 and therefore again unlikely to be *cis*-acting elements (Figure 4.3).

8	*12*	*15*	*20*
AATTTTCA TT TG GT	AATTTTCAAAAA TTA GGGG T T	GGCGACTGGCGGGAA ACTC TCT C A C A C T C	AAATTTTTTGAAAATTTTAA TTTAA CCAG TAA ACT C GT
TGAAAAAA AAG TT C	TTTTTCATTTTT CCTTA C AG C	TTTTTTTATTTTTTC CCCCCA CT T	AAAAAATGAAAAAAAAAAAA GG TCAG G GT GCTG T T T
TTTTTTTT A G	AAAAAAGAAAAA G G TAGGG	AAAAAAAAAAAAAAA GG TTG GG G GT	TTTTTTTTCTTTTTTTCCTT A CCCCTAAC CTAAA C C T C
CCTGCTCC TGGCTGT GTCT CG	GCCCGCGTGCCG CATGTTTCCTTC G CGCG	GCGGCCTCGTCACCA CGCCTGGTCAGCAGC A GTC G G G	TTTTCAAAAAAAAAAATTAA A CTGGGGG TTAATT GC GGG
TTTTTTTT A G	TTTTAATTTTTT CCTTA C A	TTTTTTTTAATTTTT AA CATTA C	CCGACAACGTGCTCGTCGCG GGATAGGGTGCGAGCGGCTC TA GCCTCAT ATT
TTTAAAAA GG CC	TTTTTCAATTTT CAGTA G	ATTTTTAAAAAAAAA TAA CGGG T	AAAAAAATAAAAAAAATTTA G TTATTG TAAAT G
TTTTTTTT G	AAAAATTAAAAA TAAGG CT	TTTTTTTTAATTTTT A CCTTA A	TTTTTTTTTTAAAAAAAAAA AAA AAAATG TTT GG
AAAAAAAA TTTTTTTT GGGGGGGG CCCCCCCC	CCGATTTGCCGG G AG T AA	AAAAAATTAAAAAAA G TAAGG T CT GC	TTCGGCGATTTGCCGATTTT CTTA A TGT G A G T
AAAAAAAA TTTTTTTT GGGGGGGG CCCCCCCC	GCCGATTTGCCG G AG G	TGCCGATTTGCCGGT A G AG TTTAAA	ATTTGTCGCGTCTAGACATG TG TCATTCGGAC TACAT GT G GG G
AAAAAAAA TTTTTTTT GGGGGGGG CCCCCCCC	GCCGATTTGCCG G AG T	GGCGACTGGCGGGAA ACTC TCT C A C A C T C	AAAAAATTAAAAAAAAAAAA G TTACGG TTTTTT GT G C

Table 4.2. The ten motifs identified using the TCM option from the upstream region of the 16 excretory cell-expressing genes. The motifs were generated using specified motif widths of 12, 15 and 20, respectively. Motifs detected in greater than one width options are highlighted in the same colour. The last three motifs in the first column (-W 8) indicate that there was no preference for any of the bases at any of the positions.

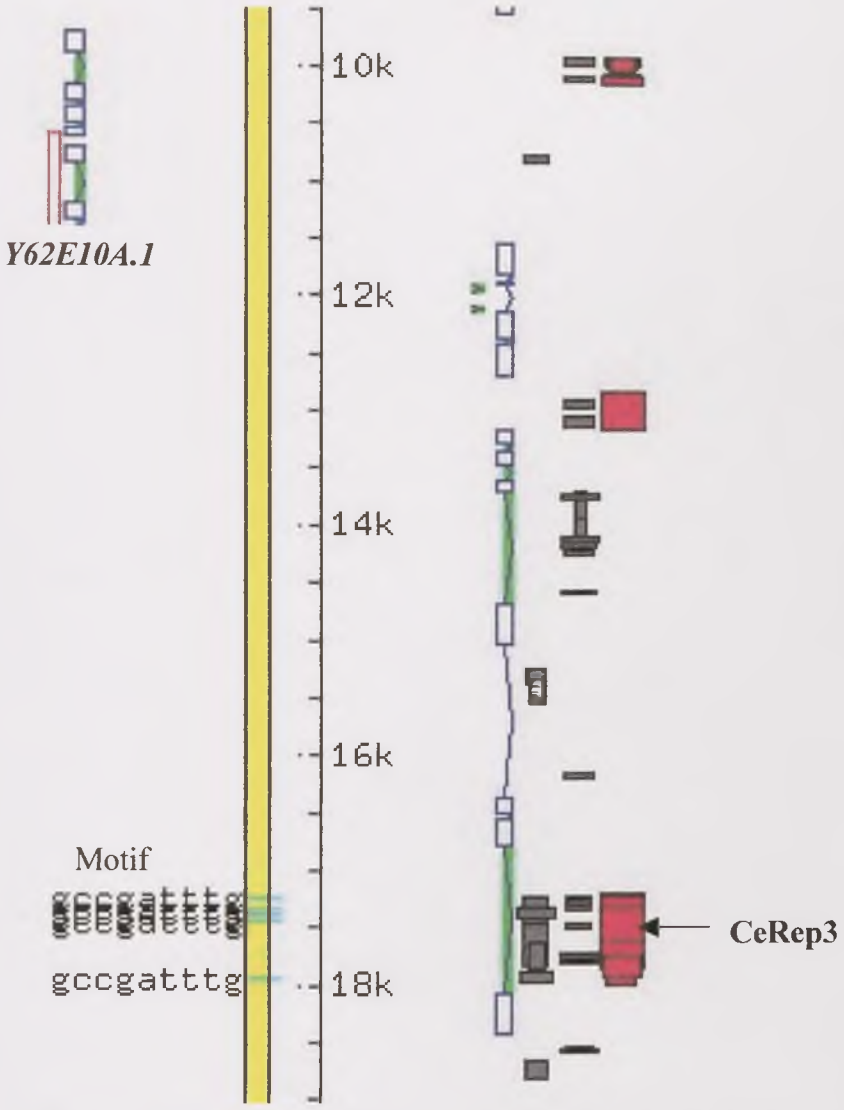

Figure 4.3. **A schematic representation of the gene structure for *Y62E10A.1* showing the motif identified to be part of the repeat element CeRep3.** The diagram shows the multiple occurrences of the motif within the repeat element indicated by the red box.

From the analyses described above it was considered that a motif width of 8 bp (or less) was too short for the MEME analyses to provide meaningful results as it identified motifs that were probably short repeats or highly non-specific due to complete degeneracy. Increasing the width to 20 bp identified motifs that also had a high AT content, but were contained in too few gene sequences. It was therefore considered that a motif width of 20 bp was too long, specificity being increased such that motifs were present in very few sequences. Therefore a motif width of 12 bp was considered to be potentially the best option for further analyses. In addition as previous studies have

demonstrated that *cis*-acting elements ranging in length from 6 bp (such as the WGATAR element involved in gut expression [Egan *et al.*, 1995], and the 7 bp vitellogenin motif TGTCAAT [Spieth *et al.*, 1985]), to 15 bp in length (as in the heat shock element aGAAtaTTCtaGAAt [Drees *et al.*, 1997]) the figure of 12 bp remained within the expected size-range for a *cis*-acting regulatory element (Sumiyama *et al.*, 2001).

Finally, the MEME software was then executed with the options –W12, -mod ZOOPS and -nmotifs stipulated as 10, 20 and 50. As expected the 10 highest-ranking motifs obtained with an –nmotifs of 20 (and 50) were identical to those obtained using an –nmotifs of 10 (Table 4.3). This option did not change the analysis but simply the number of motifs provided. Therefore it was decided that for future MEME analyses an –nmotifs of 10 should be used.

10 motifs	*11-20 motifs*	*21-50 motifs*
AATTTTCAAAAA TTA GGGG T T	GCCGATTTGCCG G AG	AAAAAAAAAAAA TTTTTTTTTTTT GGGGGGGGGGGG CCCCCCCCCCCC
TTTTTCATTTTT CCTTA C AG	GCCGATTTGCCG G AG	
AAAAAAGAAAAA G G TAGGG	CCGATTTGCCGG GT G GGT	
GCCCGCGTGCCG CGTGTTC CGAC AG GT T	CCGATTTGCCGG T G A	
TTTTAATTTTTT CCTTA C A	CCGATTTGCCGG G G	
TTTTTCAATTTT CAGTA G	CGTCTCGAGACC GCCGG C G	
AAAAATTAAAAA TAAGG CT	GGGCTCATCGAG CTT CG ACC GA T	
CCGATTTGCCGG G AG T AA	GCTCCCGTCCCG CGATT AAAA A	
GCCGATTTGCCG G AG G	TTTTTTAATTTT CCGTA G	
GCCGATTTGCCG G AG T	AAAAAAAAAAAA TTTTTTTTTTTT GGGGGGGGGGGG CCCCCCCCCCCC	

Table 4.3. Motifs generated using the options –mod ZOOPS, -W12 and –nmotifs 10, 20 or 50.

4.2.2: MEME analysis using 1 kb DNA sequences upstream from the initiation codon.

Previous studies, mainly analysing the yeast genome data, have suggested that *cis*-acting elements usually occur within the 1 kb region upstream from the first exon (Spieth *et al.*, 1985; Brazma *et al.*, 1998a; Spellman *et al.*, 1998; Gilleard *et al.*, 1997; van Helden *et al.*, 1998), although the yeast genome is more densely packed with regulatory regions tending to be more dispersed in multicellular organisms (Vilo *et al.*, 2000). Given that in the analyses described above (in which the entire upstream region to the adjacent genes were used) few motifs were identified that are potential *cis*-acting elements, the analyses was repeated with DNA sequences derived from only the 1 kb upstream region. Reducing the length of DNA analysed was expected to increase sensitivity.

While the analysis of the intergenic regions was on-going an additional gene *R13H4.5*, was shown to express in the excretory cell. Therefore for subsequent MEME analyses 17 rather than 16 genes were used as the input.

Using the 1 kb upstream region, MEME was executed with the options –W12, -nmotifs 10 and either ZOOPS or TCM (Table 4.4). As above, in the initial MEME analysis, there did not appear to be any common motifs identified by the ZOOPS or TCM options. In addition the motifs identified using the 1 kb data set were different from those identified using the entire intergenic region, with the exception of motif 2 from ZOOPS (red box, Table 4.4) which was similar to motif 10 (ZOOPS, Table 4.1). Most of the motifs identified were simple motifs and/or they were highly degenerate therefore unlikely to be *cis*-acting elements.

ZOOPS	*TCM*
TGAAAAAAAAAA G GGC G	TTTTTAATTTTT A CC A T
TCCGATTTTTCA A GCCGC	TTAATTGAAAAA AATTGACG C A A
AAAGTCAGGAAA G AAGTTC	TGCCGCCGCGCC G G TAAAACGG T GT
GCTGGCGAAACC CGA CG G TA T T A	TTTTTTTTTGAA GGGAGT
ATTTTTTTTGGA G TCT	TTTTTGGGAAAA GAAAG CT
TCGTTTTTTTCT AGCG	TTTTTGGGAAAA GTAAG
GCCGCCTCTACT G TT G CCGC G	AAAAAAAGAAAA GG TAG C G
TTTTCCAGAAAA A AG GG G G T	TTCCATCTCCCC CC TCTC TAT G
TTTCATTTCTCT C CAG C T C A	TGCGTATGGGTG G TTAT TCA G
TGGGGAAAATGA CC AC ACG T A	TATTTTTAGTTC T G T G G

Table 4.4. The ten motifs detected by MEME using the ZOOPS and TCM models from the 1 kb upstream region. Motif in red is similar to motif 2 from Table 4.1, ZOOPS column.

4.2.3: MEME analysis using the entire genomic DNA insert sequence.

The selection of DNA sequences used in the analysis were reconsidered. The precise locations of the intergenic regions and 1 kb upstream regions used in the analyses above depended on the gene annotation in WormBase. While the limits of some genes are confirmed with EST data, the precise location of other genes was solely based on GENEFINDER predictions (The *C. elegans* Genome Consortium, 1998). The ends of genes are notoriously difficult to predict as terminal exons may be omitted or aberrant exons included without the support of flanking exons and coding sequence homology. If genes are incorrectly predicted the intergenic and 1 kb upstream regions used in these analyses may not be the regions intended. In addition, gene regulatory elements need

not be located in the 1 kb upstream region or even the intergenic region as enhancers can work over considerable distances, and can even be located downstream of the coding region of a gene. However, the DNA fragments orchestrating expression in the excretory cell when fused to a reporter gene must contain the regulatory elements responsible for expression. Therefore, MEME analysis was subsequently performed using the entire DNA fragment (in the reporter gene fusion) assayed to show excretory cell expression (Table 2.2). However, use of these DNA sequences will incorporate more irrelevant information, increasing problems associated with background noise. Therefore, the analysis was initially performed with the options TCM, as this model is less sensitive to noise (Bailey and Elkan, 1994) and width 12 (for the reasons described above, Section 4.2.1).

Once more, most of the motifs detected were AT-rich (Table 4.5). For the single motif that was not AT-rich, the positions were poorly conserved and so could not be considered a candidate regulatory element. The analysis was subsequently repeated with the ZOOPS option with a negative set of DNA sequences (i.e. for genes that expressed in cells other than the excretory cell, Table 2.3). All motifs detected in the positive data set were also detected in many genes of the negative set, occasionally with a very slight variation in the base preference at one of the positions within the motif.

ZOOPS	*TCM*
GGTCTCGTCACG CGGC A T C	TTTTCAAAAATT A TC TTAA GG
AAATTTGAAAAA T	AAAAATGAAAAA G ATGGGG C
CGTGTCGAGACT GT G C A	TTTTCGTTTTTC GCTTA CT A C
CAGCAGCAGCAA TC CA	TTTTTTTTTTTT CCAAAA CC
GAAGTTGAGAGA TC AG C G	TTTAAAAAAAAA A CGCC TG
TTGCTCGTCGTC G G T TTA	GTTCTCGTCGAG CCGGCGCAGACC G CT
TTTTCAAAAAAA G	AAAATTTAAAAA TACGG C
CCGCATCTTCCG GAA AT G T G A	TTTTCCATTTTT CTAGA
TTCAATTTTTCA C A	AAAAAAAA TTTTTTTT GGGGGGGG CCCCCCCC
CCTTTCAAAATT A G G	AAAAAAAA TTTTTTTT GGGGGGGG CCCCCCCC

Table 4.5. Motifs generated by MEME using options -W12, -nmotifs 10 and both ZOOPS and TCM with the genomic DNA insert sequences of 18 excretory cell-expressing genes.

4.2.4: MEME analysis using life stage-specific genes.

A specific subset of genes (*C14A4.12*, *F10B5.1*, *F44A2.5*, *F54C9.1*, *R05G6.6*, *Y62E10A.1* and *Y70G10A.3*) were known to express at elevated levels in the excretory cell at one particular life stage, the embryonic 3-fold stage. At this stage of development, the excretory cell first differentiates, and it might be expected that a single regulatory system for driving excretory cell differentiation might be acting at this point on these genes. Therefore the entire intergenic regions of this smaller group of genes were subjected to MEME analysis. MEME was executed with the options –W 12,

-nmotifs 10 and with –mod option TCM and ZOOPS. The results obtained from the TCM model were similar to previous results, in that they were simple AT-rich motifs, and most positions contained degeneracy (Table 4.6).

ZOOPS	*TCM*
GCTCCGCCCCCG C CGTG TC	TTTTCAATTTTT C CTCT CC
TTCGTGGCGAGA GGTG	AAAAATAAACAA GCCGTA T A GG
GCGAGCTCGTCG AG T CG C	GCCCGCGTGCCG C GTTGCC TA TT G
TCGAGGCAGAGG G A T T	AAATTTTTTTTG T TA G C T G
GAAAACTGTCCG C ACG	CACAAATTTCCA GCG CGA G A A C G
CGCCACGACACC T G GG	TTAAATTTCAAA AA TG G G C C
TCCCGCCTCTAC CT G T G	GAAATGGTGACG A G CAA GAA T T
AAGAGTACTGTA T G	TTTTGTATAGAA G CCGGA
TCCGTTTGCCGG G GAG TCC C	GCATATTGTAGG GG CCC T T
CCGCCGTCCGCC G C GT T A	AAAAAAAAAAAA TTTTTTTTTTTT GGGGGGGGGGGG CCCCCCCCCCCC

Table 4.6. The ten motifs generated using the entire DNA sequences from the upstream regions of genes which expressed highly in the excretory cell at the 3-fold stage.

With the use of the ZOOPS option, the results from the analysis were unique to all previous MEME outputs (Table 4.6) with very few AT-rich motifs detected. The majority of the motifs identified were GC-rich and had a high degree of sequence complexity. In order to determine at what point motifs switched from predominantly GC- to AT-rich, the analysis was repeated but with the option –nmotifs set to 50 rather than 10. The output from this analysis revealed that all motifs were GC-rich until the motif ranked number 34, after which, motifs generated had no information content (i.e. could be either A, T, G or C at all positions) and hence were non-specific. Although

numerous motifs were identified which had complex compositions, were GC-rich and well conserved (features consistent with potential *cis*-acting elements), it is unlikely that this high number of elements would be required to regulate gene expression in a single cell type. Therefore the analysis was repeated with a negative set that consisted of the entire intergenic regions of 33 genes (the same genes as in Table 2.3) with options –W 12, -nmotifs 34, -mod ZOOPS. The negative set consisted of genes that showed expression in cells other than the excretory cell. The results from this analysis demonstrated that all the motifs identified in the positive set were also detected in many of the genes in the negative set. In addition, it is likely that due to the fewer number of genes used in the positive set several complex motifs were identified by chance.

4.2.5: MEME analysis using genes that express exclusively in the excretory cell.

The majority of genes that expressed in the excretory cell also expressed in other cell types. However four genes (*B0285.6*, *T20B5.3*, *Y70G10A.3* and *Y113G7B.24*) were known to express solely in the excretory cell. Therefore to potentially improve the likelihood of detecting *cis*-acting elements, MEME analysis was repeated using the 1 kb upstream region of only these four excretory-cell specific genes. The analysis was executed with the options –W 12 and –nmotifs 10 with both TCM and ZOOPS. As observed previously, the majority of motifs detected with the TCM model were short runs of As or Ts or a mixture of As and Ts and were therefore unlikely to be *cis*-acting elements (Table 4.7). There was no overlap in the motifs detected using either the ZOOPS or TCM models. The ZOOPS model reported a few motifs that had a complex nucleotide arrangement and therefore were more likely to be *cis*-acting elements than simple repeats, however, they were highly degenerate at the majority of positions. Based on this analysis it was possible to conclude that analysing only the four genes which expressed solely in the excretory cell did not generate more likely *cis*-acting elements than in the preceding MEME analyses.

ZOOPS	*TCM*
TCTCTTCTCGCC G ACGCG AA G	AAAATTTTAAAA A CGCG CT
ACCTTTCCTCCC CATA GGGT C G	TTCAATTTTTC CAATTG T ACG
GGGGGTTCGAAC CCCA GCGA G T	AAAAAAAAAAAA G GGGG G C
CCCAGTTGCACA G TTC CACG	CCTTTTCTCCCC AC CGTTA CA G
TTTCGTGGGAAC AT GCA A C G	TCATTTTTTTGG CAGA AAGA
GGAAAAAAAATG A GG C	TGCCTGTTCGAG CCGGG GGA A A TC C
TTTTGCGCTAGG TA TAT G T	ACTGATTGCTCT TG C G AGC A
AAAATCGATGAA C G CT C	TTTTAAAATTTT G C TC C T C
TTCACTTTTTAG CA G T C G	CAGGCCACAAA T CCAA GT G
GGGCTTGCTGGA CAAG A A TG T T G	GTGGCCGGAGT A AAA T GT T

Table 4.7. The ten motifs identified using the 1kb DNA sequences from the upstream regions of genes that solely express in the excretory cell.

4.2.6: MEME analysis using a larger excretory cell data set.

An additional 23 genes which express in the excretory cell were then incorporated into the analysis. These additional genes had been identified from the WormBase database and the published literature (Section 3.4.2). The 1 kb DNA sequences of these 23 genes (Table 2.9), in addition to the now 20 genes identified in this laboratory by the reporter gene fusion approach, were used to execute MEME with options –W 12, –nmotifs 10 with both ZOOPS and TCM. Again, this data set generated many AT-rich motifs which had low sequence complexity (Table 4.8). The highest-ranking motif from the ZOOPS set was similar to motif 4 in the TCM set. However, this motif was poorly conserved with 9/12 ambiguous positions.

ZOOPS	*TCM*
GCCGCCGCCGAC GGTT T CCA G	TCATTTTTTTTT AA A CGAA T CC G
GAAAAAGAGAA A CGA A G	AAAAAAGAAAAA G TCGGGGG
TTTTTTTCATTT CCG	TTTTTTAAAAAA AAA CGG T G
AAAAATAAAAAA A G G G	CCGCCGCCGCCG GGCT T TCG C T A
TCTCCGTTTTTC TGG AGC C T	TTTTTTATTTTT ATAA G
AAAAATTGAAAA GG G C GG	TTTTGTGGAAAA GTGCAG A T
TTTTAATTTTTC CC AA	AAAAAAAAAAAA T TTT
GGAAGAAAAACG T GCAG TC	TTTTTTTGGAAA GAGTAG G C
GGAAGAAAAACG CTT A GC T G	AAAAAAAAAAAA TTTTTTTTTTTT GGGGGGGGGGGG CCCCCCCCCCCC
TCACCTCCACTC C CG CT TAC T G	AAAAAAAAAAAA TTTTTTTTTTTT GGGGGGGGGGGG CCCCCCCCCCCC

Table 4.8. The ten motifs generated using the 1 kb DNA sequences from the upstream regions of 43 genes expressing in the excretory cell. The red underscore identifies similarities in motifs generated by the ZOOPS and TCM models. Motifs highlighted in orange were very similar.

4.3: Discussion.

A computational strategy using the MEME software has been used to identify candidate *cis*-acting regulatory elements responsible for gene regulation in the excretory cell of *C. elegans*. This software was originally developed by Bailey and Elkan (1994) to identify motifs that are common to a given set of DNA sequences. In previous bacteria and yeast studies, MEME has been able to detect motifs from a set of DNA sequences without any *a priori* knowledge of either the frequency of occurrence, or the precise location of the motif (Bailey and Elkan, 1994; Workman and Stormo, 2000). Furthermore, the output from MEME allows for degeneracy within a motif. This was considered to be particularly important because *cis*-acting elements are generally poorly conserved (Bussemaker *et al.*, 2001). In addition, MEME has been shown to detect motifs in bacteria and yeast even if as little as 20% of sequences in the data set contain the particular element (Bailey and Elkan, 1994).

A number of experiments were performed using several data sets with varying -mod (motif distribution models), –W (motif length), and –nmotifs (number of motifs required for output options). Initially the experiments were performed using the entire intergenic regions of all genes in the data set with all three frequency models OOPS (one occurrence per sequence), ZOOPS (zero or one occurrence per sequence) and TCM (any number of occurrences per sequence). There was considerable overlap between the OOPS and ZOOPS results whereas the TCM output was quite distinct. Since the potential element was not expected to occur in every sequence in the positive set the OOPS (one occurrence per sequence) option was not considered appropriate for further analyses. The TCM option was considered the best as it allowed for multiple occurrences of the motif within a sequence. This was an important factor in *cis*-acting element detection because previous studies have shown that *cis*-acting elements can occur multiple times within a sequence (Spieth *et al.*, 1985; Gaudet and Mango, 2002). By contrast it was also considered possible that the motif may not be present in all sequences as some genes may use a different mechanism for directing excretory cell expression (Vilo *et al.*, 2000). However although the TCM option had the apparent advantage of detecting multiple motif occurrences within a sequence this also proved to be a disadvantage as this model was particularly susceptible to detecting unwanted simple tandem repeats which were unlikely to be candidate *cis*-acting elements.

Although the ZOOPS option identified motifs with high sequence complexity, most of the positions within these motifs were poorly conserved and therefore considered unlikely to represent *cis*-acting elements. In addition, the majority of motifs from the ZOOPS output were demonstrated to be part of complex repeats present in various genes and therefore again unlikely to be *cis*-acting elements.

With the majority of the different analyses, MEME largely detected simple AT-rich motifs. These motifs were considered non-specific as the *C. elegans* genome, particularly the intergenic regions are AT-rich (Kent and Zahler, 2000). The motifs that were GC-rich or complex in sequence were very poorly conserved and were also detected in a negative set of genes. Most of the motifs detected by MEME that were not simple runs of As and Ts were found to be part of repeat elements. This was best exemplified with the output from the TCM option where some of the motifs identified occurred in a single gene with a frequency of greater than 20.

Based on the results obtained from the numerous experiments performed with MEME, it is possible to conclude that its usefulness for detecting *cis*-acting elements in *C. elegans* is limited. This is possibly because repetitive DNA is a major component of higher eukaryote organisms (such as *C. elegans*) and also that the regulatory regions are widely dispersed. As has been reported previously (Bailey and Elkan, 1994) MEME may be more suited to simpler organisms (such as yeast and bacteria) which are known to have a more compact genome and the promoter regions of genes are situated close to the gene translation start sites, predominantly within 600 bp upstream of the gene (Palin *et al.*, 2002; Workman and Stormo, 2000).

Ultimately, as the output from MEME was vast and complicated and the software was extremely difficult to modify, it was considered necessary to obtain different software to identify *cis*-acting elements. This software was SPEXS and will be described in detail in the next chapter.

Chapter 5

Chapter 5: The use of SPEXS software to identify *cis*-acting elements from the promoter regions of co-regulated *C. elegans* genes.

5.1: Introduction.

As discussed in the previous chapter MEME failed to identify any candidate *cis*-acting elements which direct gene expression in the excretory cell of *C. elegans*. Therefore a different detection strategy was adopted using software known as SPEXS (Sequence Pattern EXhaustive Search). The SPEXS software is based on a conceptually simple, pattern-generation algorithm that was developed for the analysis of DNA sequences on a genomic scale by performing an enumerative (also known as exhaustive) search for all patterns (motifs) up to a certain length (Brazma *et al.*, 1998b). SPEXS has previously been successfully used to detect *cis*-acting regulatory elements from a set of co-regulated genes in the yeast genome (Brazma and Vilo, 2000).

The input to the SPEXS algorithm is a set of unaligned DNA sequences, which are used to construct the data structure of a suffix tree (Giegrich and Kurtz, 1995; Giegrich and Kurtz, 1997). It is constructed such that, at each node the pattern length is extended by a single nucleotide (Figure 5.1). The suffix tree follows similar rules to that applied to inheritance, where the "child" (a node) inherits the pattern associated with its "parent". At each node of the suffix tree, an occurrence list is maintained (the only computational calculation involved in the method) giving all the positions in the input sequence that match the particular pattern. In order to expedite processing the "children" are only generated if the pattern of the "parent" is detected in the input sequences. The SPEXS output contains a list of the particular patterns identified in the input sequences, the total number of occurrences of the pattern, and the number of sequences that contain the pattern in question.

In an effort to conserve memory usage and increase the rate of processing, the SPEXS output contains patterns no greater than 20 bp in length and identifies motifs that are only present above a particular frequency (specified by the user). In addition the speed of the SPEXS algorithm is increased by a so-called "pruning" strategy (Brazma *et al.*, 1998c).

This involves searching for patterns within the suffix tree in a "depth-first" manner. For example if the pattern to be searched for is TGA, only prior knowledge of the presence of the pattern TG is required rather than TG and GA which is the case with the alternative "breadth-first" search. The breadth first approach involves searching for a pattern across the tree and thus requires the presence of both patterns TG and GA before confirming the presence of the pattern TGA (Figure 5.1, solid lines). Although, the breadth-first approach is less time-consuming, the search is restricted to very short motifs, because the number of motifs to be stored in memory increases rapidly. Therefore, the depth-first search is used by most enumerative algorithms including SPEXS (Sagot *et al.*, 1997, Neuwald and Green, 1994; Jonassen *et al.*, 1995).

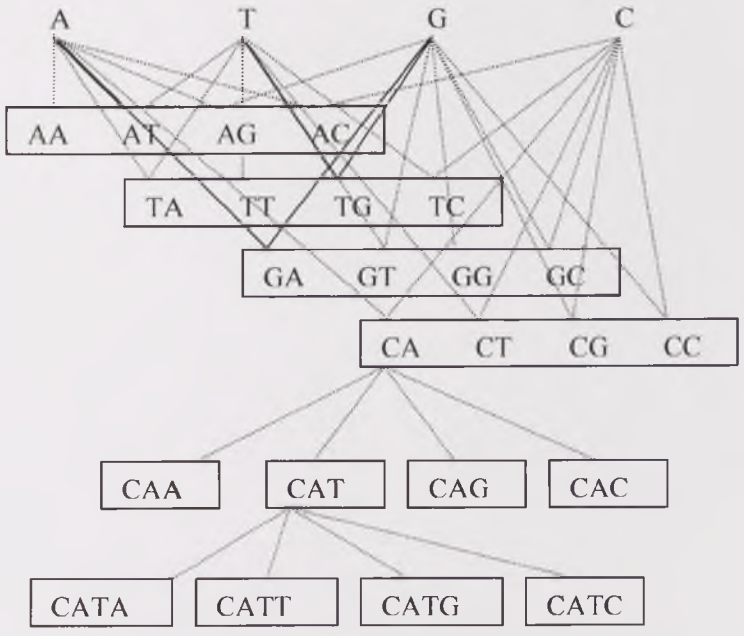

Figure 5.1. A schematic representation of the generation of a suffix tree employed in motif identification. The pattern TGA is indicated by the presence of TG and GA, solid lines.

5.1.1: Improvements to the SPEXS output.

Although SPEXS has previously been used in the identification of *cis*-acting elements in yeast there are a number of problems inherent to the output provided by this software. One significant problem associated with SPEXS is that it assumes that the forward and reverse

complements of a particular motif are two separate motifs, thereby doubling the number of motifs in the output (Giegerich and Kurtz, 1995). Therefore a computational strategy will be implemented, such that forward and reverse complements are reported as a single motif.

Perhaps the most significant problem associated with SPEXS is that the output is generated on an exhaustive basis without taking into consideration biological variables that are characteristic of *cis*-acting elements. As a consequence thousands of patterns are identified, the majority of which are highly unlikely to be *cis*-acting elements. To overcome this problem, a screening protocol will be devised which will be based on defining particular thresholds for variables that are likely to be important in predicting *cis*-acting elements. These potentially important variables are; the precise location of each motif within the sequence (intergenic, intronic or exonic), the frequency of occurrence in positive versus negative data sets and the DNA sequence complexity. Formulas will be developed to calculate the individual scores for each of the three factors. Each of the three factors will be assigned different weightings depending on the perceived importance of each factor in detecting potential *cis*-acting elements. The sum of the three individual scores will provide the Total Score for each of the motifs in the list generated by SPEXS. The motif with the highest Total Score will be considered the most likely *cis*-acting element. This strategy will allow motifs, which most fulfill the anticipated characteristics of a candidate *cis*-acting element to have the greatest score. For example, a motif occurring with a high frequency close (upstream) to the translational start, and with high DNA sequence complexity, will be given the greater score.

A final major caveat associated with SPEXS is that it does not allow for degeneracy and only identifies patterns that are identical to the input sequences. Therefore the use of a novel strategy to introduce degeneracy in the identification of motifs will also be employed.

The application of the scoring strategy and the allowance for degeneracy in the SPEXS output will be initially applied to the genomic DNA fragments of genes that had been used in reporter gene fusion constructs. These genomic DNA fragments will be used because these sequences are known to drive expression in the excretory cell of *C. elegans* and therefore they must contain the appropriate *cis*-acting regulatory elements.

5.2: Results.

At the commencement of the SPEXS analysis, the positive set of excretory cell expressing genes consisted of 18 genes (Table 2.2). Of these 18 only two genes, *Y70G10A.23* and *Y113G7B.24* expressed exclusively in the excretory cell with the remainder expressing in cell groups in addition to the excretory cell. All gene expression patterns were determined in the laboratory of Dr I. A. Hope using the reporter gene fusion technique [http://129.11.204.86:591]. To increase the specificity of *cis*-acting element detection a negative set (Table 2.3) consisting of 33 DNA sequences, (representing genes not expressed in the excretory cell, but expressing in other cell types) was used to affect a comparison. The forward and the reverse complements of both data sets were used to identify potential *cis*-acting regulatory motifs that may be present in either forward or reverse orientations.

The output from the SPEXS analysis consisted of a motif followed by the number of sequences (genes) containing the motif in the positive and negative sets, respectively (Figure 2.4). The SPEXS software analysis enumerated thousands of motifs ranging in length from 1 bp to 20 bp, however a large proportion of the motifs could be dismissed immediately because they occurred more frequently in the negative set, or because they were between one and four bp in length (and therefore highly unlikely to be *cis*-acting elements). After eliminating these motifs, the number of remaining motifs was still too great for experimental verification (Section 8.2.1). Therefore, the scoring strategy based on motif frequency, location and sequence complexity was devised and implemented to identify the most likely candidate *cis*-acting element.

5.2.1: Motif frequency.

For a motif to be considered over-represented in the positive set a threshold for the difference in frequency of occurrence between the positive and negative gene sets had to be decided upon. With at least three more occurrences in the positive than in the negative set, the number of motifs detected was far too great for further analysis. When the difference in frequency between the positive and negative set was increased to at least four the number of motifs detected was approximately 600 (Figure 5.2).

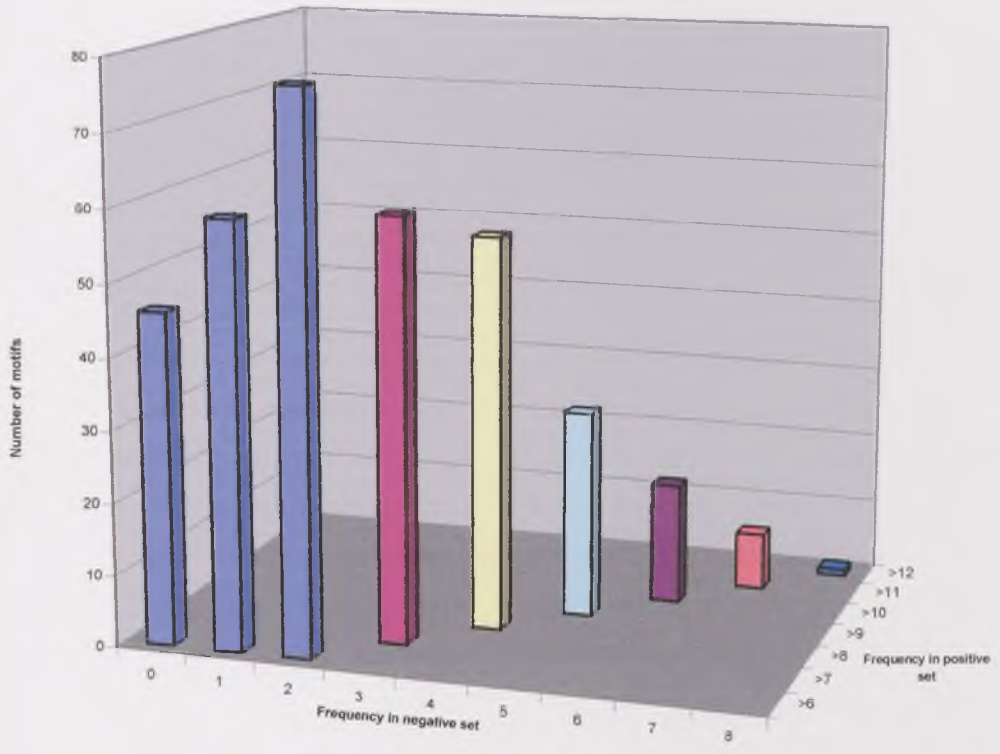

Figure 5.2. Frequency distribution observed for motifs obtained from the genomic insert approach. The graph shows the number of motifs detected using the cut-off frequencies defined for the positive and negative sets.

The SPEXS software identified a motif occurring in both reverse and forward orientations as two separate motifs thereby increasing, by a factor of two, the number of candidate *cis*-acting elements. To remove this redundancy before further analysis, an algorithm was designed to identify and eliminate all reverse-orientation duplicates from the motifs satisfying the frequency threshold. This reduced the number of motifs to be considered for further analysis to 351, (Figure 5.2), slightly over half of the number of motifs initially identified, due to the occurrences of palindromic motifs. In addition, the presence of these palindromic motifs in both orientations doubled the frequency of occurrences which therefore had to be halved manually.

The frequency of occurrence for each of the motifs was obtained from both the positive and the negative sets. This data was initially obtained directly from the SPEXS output,

however, random manual checks of the frequency data demonstrated that very occasionally these values were not always accurate and required additional confirmation.

Therefore a program was developed to determine the frequency of occurrence in both the positive and the negative sets (Section 2.5), the exact location of each motif within a DNA sequence, and the identity of the genes that contain the particular motif (information not provided by SPEXS). Once the Fragment Frequency (number of sequences containing the motif) and Total Frequency of occurrences (includes multiple occurrences within a sequence) were ascertained, they were used to calculate a Frequency Value, using the formula shown below.

Frequency Value calculations:

Frequency Value = $(2 \times a) + b - (2 \times c) - d$

a = number of sequences in positive set in which a motif is detected
b = total number of occurrences of the motif in the positive set
c = number of sequences in negative set in which a motif is detected
d = total number of occurrences of the motif in the negative set

This formula provides a higher weighting to the first occurrence of a motif in the positive sequence, and a reduced weighting to any subsequent occurrences of the motif within the same sequence (Figure 5.3). This reduces the problem of assigning repeat elements (usually present with multiple copies within a sequence), which are unlikely to be *cis*-acting elements, a high frequency value.

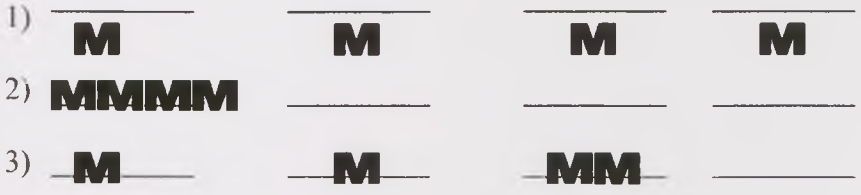

Figure 5.3. Possible distributions of motifs within DNA sequences.
Applying the frequency formula, scenario 1 has the greatest score; scenario 3 has an intermediate score and scenario 2 has the lowest score. Each line represents a DNA sequence and "**M**" represents the occurrence of a motif within a sequence.

5.2.2: Motif position.

Studies employing promoter deletion approaches have demonstrated that *cis*-acting elements usually occur in upstream regions or occasionally in the introns of genes (Haerry and Gehring, 1996; Harfe and Fire, 1998; Zhang and Emmons, 2000). Therefore it was important that the scoring system recognised the importance of the precise location of the motif in terms of intergenic, intron and exon regions. The intergenic category was sub-divided into two additional categories; less than 1 kb or greater than 1 kb from the translational start site.

Position Value = R x frequency of occurrence in region X

R = weighting assigned to region X (*i.e.* intergenic, intron and exon)

The R values are described in detail in Section 5.2.5.

5.2.3: Motif complexity.

The DNA sequence complexity (Wootton and Federhen, 1993) was calculated using the formula defined by Wootton (1994) shown below (discussed in further detail in Section 2.7).

$$Complexity = -\sum_{i=1}^{4} Pi \ln Pi$$

An analysis of complexity revealed that the vast majority of the motifs occurred with a value of 0.60 or greater (Figure 5.4). A small number of motifs (50 of 351) occurred with a value of 0.60 or below. Typical examples of motifs with the lowest complexity value (0.29) included "TTTTTTTTTTTG", the motif with the greatest Complexity Value (1.39) was "TTCGACGA".

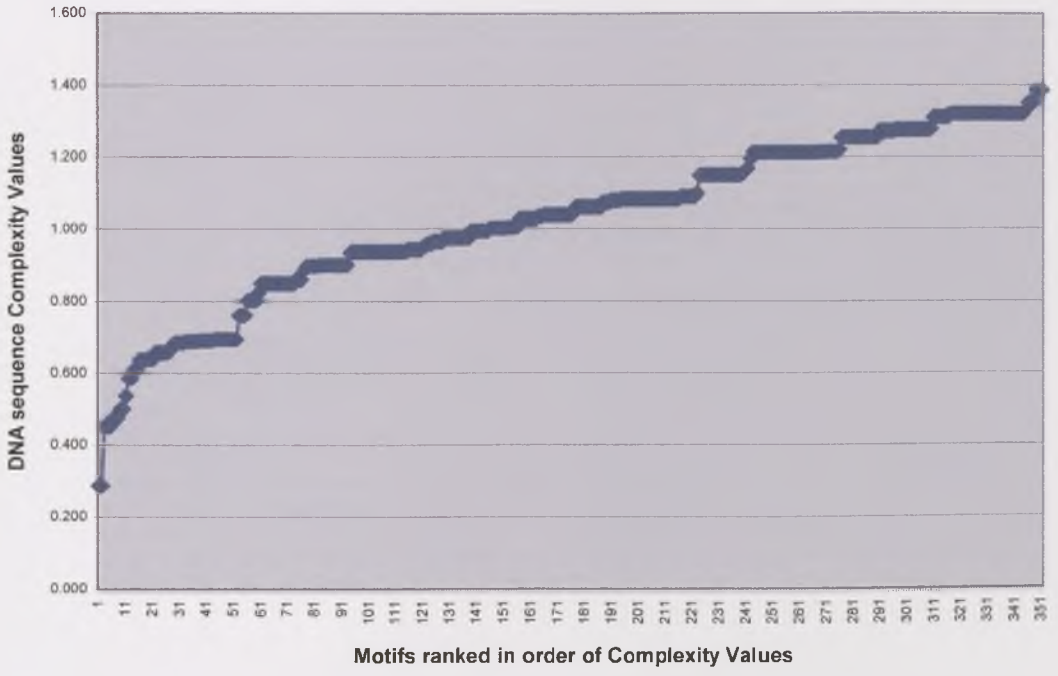

Figure 5.4. **DNA sequence complexity value distribution of motifs obtained from the genomic insert approach.**

Although it was clear that frequency, location and complexity are important aspects of a *cis*-acting element, the respective weighting that each of these variables should be assigned was more difficult. Ultimately, the degree of weighting assigned to the three factors was based on previous studies analyzing *cis*-acting elements in the yeast genome (Spellman *et al.*, 1998, Brazma *et al.*, 1998a, Hughes *et al.*, 2000; van Helden *et al.*, 1998) and promoter deletion experiments for *C. elegans* (Gilleard *et al.*, 1997; Harfe and Fire, 1998; Culetto *et al.*, 1999; Zhang and Emmons, 2000). It was decided that frequency and position were to be given equal weightings (each of 45%) representing the likely crucial nature of these variables in the identification of the potential *cis*-acting element. DNA sequence complexity was not considered to be as important as the other two variables and therefore was given a lower weighting of 10%.

5.2.4: Generation of a score based on frequency of motifs.

The frequency score was calculated using the formula:

$$\textbf{Frequency Score} = 45\% \times \frac{\text{(Frequency Value of specific motif– lowest Frequency Value)}}{\text{(highest Frequency Value – lowest Frequency Value)}}$$

This formula also linearises the Frequency Scores from 0 to 45%. Close examination of the Frequency Score demonstrated that approximately half of the motifs scored between 20 and 25% whilst two motifs had the highest scores of between 40 and 45% (Appendix II).

5.2.5: Generation of a score based on position of motifs.

The highest score (45%) was assigned to motifs that were found within 1 kb upstream of the translational start site. Motifs located beyond the 1 kb region but remained within the intergenic region scored 35%. Motifs occurring within the first intron were assigned a score of 30%. Motifs located in exons, and therefore most unlikely to be candidate *cis*-acting element, were assigned a negative score -5%. The calculations were performed as followed:

$$\textbf{Position Score} = \frac{\text{Sum of all Position Values of a motif}}{\text{Total number of occurrences of a motif}}$$

An additional further emphasis on frequency was removed by dividing the sum of the Position Values (Section 5.2.2) by the Total Frequency (Section 5.2.1). In addition this formula also linearises the Position Score from 0 to 45%. Examination of the Position Scores demonstrated that approximately 50% of motifs had a score of between 25 and 30% (Appendix II).

5.2.6: Generation of a score based on DNA sequence complexity.

The Complexity Scores were calculated using the following formula:

$$\textbf{Complexity Score} = 10\% \times \frac{\text{(Complexity Value of specific motif – lowest Complexity Value)}}{\text{(highest Complexity Value – lowest Complexity Value)}}$$

The score for DNA sequence complexity was anticipated to be the least important variable in predicting a potential *cis*-acting element. The Complexity Score demonstrated a similar trend as the Complexity Value (Figure 5.2.4) and the vast majority of the motifs had a score of 5% and above.

5.2.7: Evaluation of the Total Score.

The Total Score for each of the motifs was obtained by summing the individual scores for motif frequencies, positions and complexities. When the motifs were ranked in order of Total Scores there was a 6% difference between the patterns with the first and second – ranked scores. For the remainder of the identified patterns there was a gradual, incremental decrease in the Total Score (Figure 5.5).

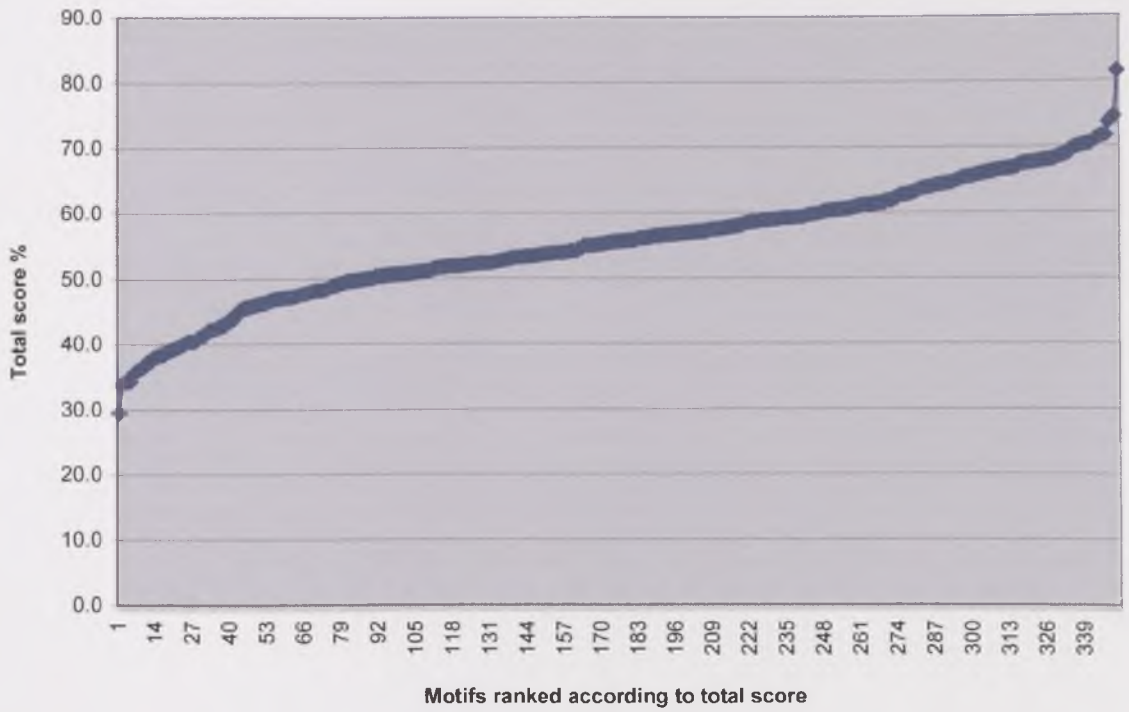

Figure 5.5. **Distribution of the Total Score obtained from the genomic insert approach.**

The highest ranking motif was 8 bp in length, high in sequence complexity and occurred with a frequency of six in the >1 kb upstream region, and four in < 1kb region. It was not present in intron or exon regions (Table 5.1). It occurred with a frequency of two in the

negative set. The second highest-scoring motif also contained characteristics of a potential *cis*-acting element. It occurred three times more often in the < 1kb upstream region than the > 1 kb upstream region and it also occurred with a frequency of two in the negative set. However, it obtained a lower score because of the lower overall frequency (Table 5.1). The third ranked motif was considered unlikely to be a true motif as it had a very low sequence complexity and only possessed a high Total Score (almost as high as the second-ranked motif) because of the high frequency of occurrence. In addition, this motif was evenly distributed between the < 1kb upstream region, introns and exons. The fourth ranking motif was more interesting in that it had a moderately high sequence complexity and it only occurred beyond the 1 kb upstream region except for single occurrences in an intron and an exon. It was difficult to assess how significant this result was as it was possible that these occurrences could have been just beyond the 1 kb region rather than more widely dispersed in the intergenic region. These four motifs were used to search the positive set in order to identify the genes containing these motifs and to locate the exact positions of each occurrence of the motifs. This analysis demonstrated that the fourth motif did not occur close to the 1 kb upstream region but was widely dispersed suggesting that the motif was less likely to be a *cis*-acting element. Appendix II shows the Total Scores calculated for each of the motifs meeting the frequency threshold criteria defined in Section 5.2.1

		Positive set		Negative set										
Motif_rank	Motif	Frag.Freq	Tot.Freq	Frag.Freq	Tot.Freq	Freq_score	Comp.	Comp_score	> 1kb	< 1kb	Tot.introns	Exon	Pos_score	TOTAL_SCORE
1	ATCGATCA	10	11	2	2	33.39	1.32	9.41	7	4	0	0	38.64	81.43
2	AGTATACT	8	8	2	2	23.23	1.26	8.81	2	6	0	0	42.50	74.54
3	AAAAAAATTAA	13	16	3	3	45.00	0.47	1.70	6	4	3	3	27.50	74.20
4	GTATCAAA	11	13	3	3	34.84	1.21	8.43	11	0	1	1	30.39	73.65
5	TTATTCAAT	11	12	4	4	29.03	0.94	5.91	5	6	0	1	36.67	71.61
6	CGATCAGT	6	6	0	0	23.23	1.39	10.00	4	2	0	0	38.33	71.56
7	GCCTAATTT	6	7	0	0	24.68	1.27	8.97	5	2	0	0	37.86	71.51
8	TAATATTGA	8	8	1	1	27.58	0.96	6.17	6	2	0	0	37.50	71.25
9	TAATAACT	11	14	4	4	31.94	0.97	6.25	7	5	0	2	32.86	71.05
10	AAATTGAATA	11	12	3	3	33.39	0.90	5.56	8	2	1	1	31.67	70.61

Table 5.1. **The top ten ranking motifs obtained with the genomic insert approach.**

5.2.8: Incorporating a score for degeneracy.

It is well-recognised that *cis*-acting elements may contain several positions of degeneracy. It was predicted that modifying the SPEXS output to incorporate a score for the allowance of degeneracy at a position within the motif would further distinguish and identify motifs that are the more promising candidates for *cis*-acting elements.

As expected the output generated by the SPEXS analysis contained numerous examples of motifs which were identical other than a single base difference in length (Example 1) or containing a single base discrepancy at an internal position within the motif (Example 2).

Example 1

```
Motif A   5'ATATATATAT 3'
Motif B     ATATATATAT A
```

Final Score of Motif A = Total Score of A + (1/4 x Total Score of B)

Final Score of Motif B = Total Score of B + (1/3 x Total Score of A)

Example 2

```
Motif C   5' ATATATATAT 3'
Motif D      ATACATATAT
```

Final Score of Motif C= Total Score of C + (1/3 x Total Score of D)
Final Score of Motif D= Total Score of D + (1/3 x Total Score of C)

Considering Example 1, to account for degeneracy the score of the shorter motif (Motif A) was increased by 25% of the value of the Total Score of Motif B because there were potentially four longer motifs that could contribute in this way for each shorter motif. The longer motif (Motif B) had greater specificity at the 3' end (with an A nucleotide) and therefore the Total Score of B was increased by 33% of the Total Score of Motif A because all instances of Motif A are followed by one of three bases. (If Motif A is followed by the fourth base it will have been counted as motif B.)

This strategy was based on the premise that if motifs identified were similar then they were more likely to represent variations of the same *cis*-acting element and as a consequence they warranted a higher Final Score. Based on this strategy it was considered that factoring in degeneracy into the Total Score, may lead to a system in which several similar motifs ultimately contribute to the score of a single motif (Figure 5.6). The inclusion of a degeneracy score was limited to one position within a motif because degeneracy at two or more positions would have resulted in a large number of sequence permutations of a motif.

```
Motif A          ATCTGTCTCT
Motif B          ATCTGTCTCTA
Motif C          ATCTGTCTCTAT
Motif D        AATCTGTCTCT
```

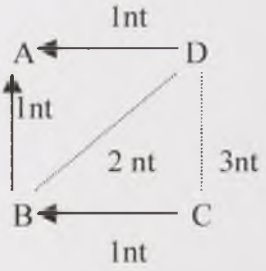

Figure 5.6. Schematic representation of the mechanism by which several motifs can contribute to the Total Score of a single motif. The score of Motif C contributes to the score of Motif B as there is only a difference of a single nucleotide between the two motifs. In turn the score of Motif B (together with D) contributes to the score of Motif A.

A single nucleotide mismatch at an internal position within the motif was accounted for using a strategy shown in Example 2. A value of 33% of the Total Score of one motif was added to the other motif, and vice versa, because, for each motif, there are potentially three other motifs that could contribute to the score when allowing for variation in that position. This would enhance the rankings of both motifs as they were variations of the same motif.

Once the Final Scores had been determined the motifs were re-ranked (now incorporating the allowance of degeneracy). This altered the ranking of the motifs considerably and although the highest scoring motif remained at the same position, the subsequent nine best

motifs of the original ranking were reduced to a lower rank (Table 5.2). The highest scoring motif ATCGATCA was detected within the 1 kb upstream regions of genes *R13H4.5*, *B0285.6*, *F44A2.5* and *Y62E10A.1* (Figure 5.7). Examination of the expression pattern descriptions of these genes revealed that all except gene *B0285.6* had one other component of expression (the pharynx) in common in addition to the excretory cell. The gene *B0285.6* expressed solely in the excretory cell. In addition, this motif was found within the 2 kb upstream region of three further genes, *C46C2.1*, *R12C12.6* and *Y18D10A.23*. The highest scoring motif ATCGATCA had one other variation of the motif, ATCGATCT, contributing to the Final Score. The motif ATCGATCT was present with six occurrences in the upstream region. Three of the six occurrences were in the 2 kb upstream region in the genes, *C46C2.1*, *F54C9.7* and *T20B5.3* (Figure 5.7).

		Positive set		Negative set									
Motif_rank	Motif	Frag.Freq	Tot.Freq	Frag.Freq	Tot.Freq	Complexity	> 1kb	< 1kb	Tot.introns	Exon	Total score	Original rank	Final score
1	ATCGATCA	10	11	2	2	1.32	7	4	0	0	81.43	1	122.22
2	AAAATTAATTT	8	11	2	2	0.69	0	7	1	3	61.24	81	120.22
3	ATATCAAG	7	8	2	2	1.21	5	1	0	2	55.00	186	119.35
4	AAAAATCGGA	6	8	1	1	1.09	6	2	0	0	66.57	42	108.98
5	ATATGTGA	8	10	5	5	1.08	8	1	0	1	52.30	223	102.29
6	AATCATTAT	7	7	2	2	0.86	4	2	1	0	51.83	233	102.16
7	AAAAAATTTTTT	6	14	0	0	0.69	2	6	3	3	61.45	47	98.33
8	AAAATCGGA	8	10	3	3	1.15	7	2	0	1	62.62	77	97.32
9	AAATTTTTAT	7	10	4	5	0.67	6	1	0	3	49.91	328	96.98
10	AAAATCGTA	9	10	4	4	1.15	4	3	0	3	54.17	193	95.95

Table 5.2. **The top ten ranking motifs obtained with the genomic insert approach factoring in degeneracy at one position within the motif.**

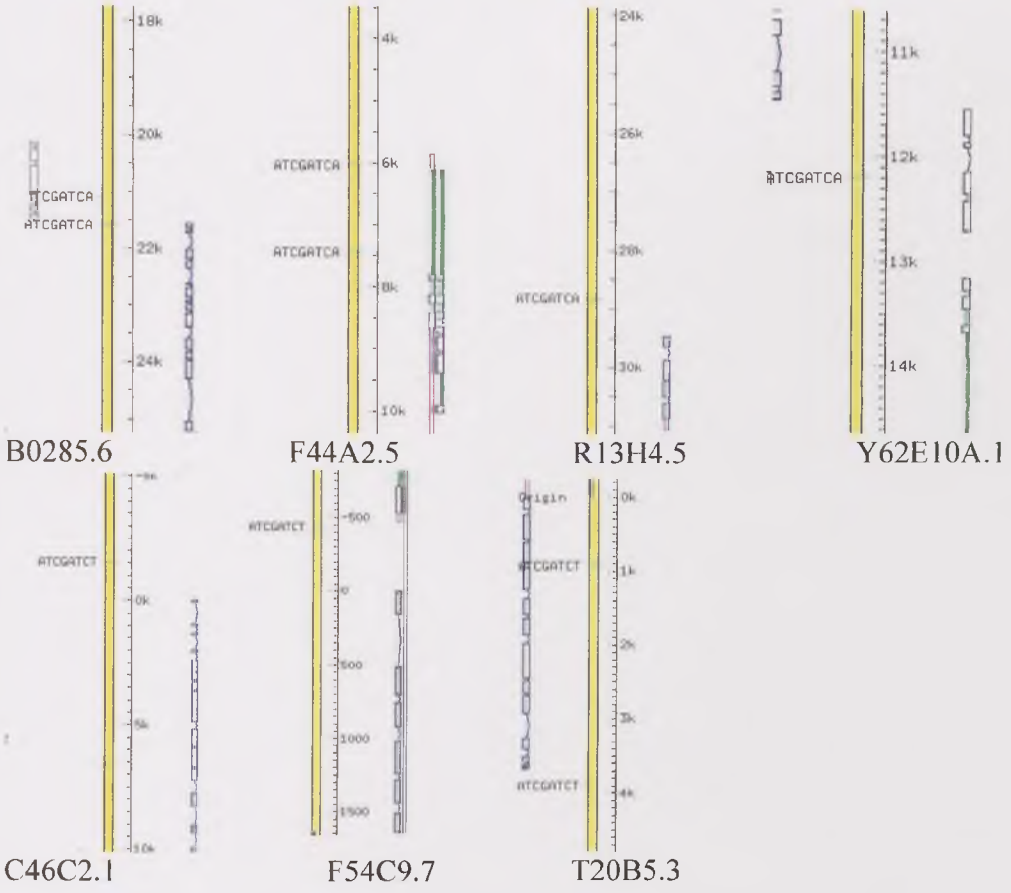

Figure 5.7. A schematic representation of genes containing the elements ATCGATCA and ATCGATCT within their 1 kb upstream regions. For each panel, the gene shown to express in the excretory cell is to the right of the yellow bar (except for genes T20B5.1 and Y62E10Aa.1), transcription orientation being down the page and running out of the view shown.

5.3: Discussion.

In this chapter the use of SPEXS software to identify potential *cis*-acting regulatory elements that drive gene expression in the excretory cell of *C. elegans* has been described. Although SPEXS has previously been used to identify *cis*-acting elements in yeast (Brazma and Vilo, 2000) there were a number of caveats inherent to the software that needed to be overcome for its use in a more complex organism such as *C. elegans* (Section 5.1).

A major problem associated with SPEXS can be attributed to the principle by which the software identifies motifs. SPEXS identifies motifs in an exhaustive manner, simply based on the sequence composition and frequency of occurrence, without considering any biological factors. Therefore a vast number of patterns are identified that are unlikely to be *cis*-acting elements. To overcome this problem a scoring strategy was designed and implemented which assigned a high score to motifs that, based on known biological features, are more likely to be *cis*-acting elements.

A negative set of sequences were used in the analysis to identify and eliminate the motifs occurring by chance or involved in the general regulation of genes (such as housekeeping genes). In addition, the genes in the negative set were selected on the basis that they showed expression in cell groups other than the excretory cell so as to identify and eliminate other motifs that may be involved in the regulation of additional components of expression observed with genes in the positive set. The negative set was made as large as possible (33 genes), until software limitations prevailed, so as to include genes showing expression in a wide range of cell groups to thereby improve the sensitivity of the scoring strategy.

The genomic DNA insert sequences (approximately 5 kb in length) of the 18 genes in the positive set were used as input to the SPEXS software because it was known with certainty that the motif was present in the data set. However, a consequence of using such large fragments was that a larger number of non-specific motifs were identified and therefore the analysis became less sensitive. This was demonstrated by the detection of 45 different motifs using a frequency threshold of zero in the negative set and six in the positive set.

This number of motifs was considered too great as it was very unlikely that 45 motifs would be involved in regulation of gene expression of a single cell.

The SPEXS software can not distinguish between sense and antisense strands of DNA. Therefore to maximize the possibility of identifying a motif, both the forward and reverse complements of all sequences were analysed and as a consequence, all identified motifs were present in both orientations. Furthermore all palindromic motifs had their frequency of occurrences artificially doubled as they were present, of course, in both orientations of the gene.

It was anticipated that the ranking of the Total Scores for each of the motifs would discriminate between a small number of patterns that are likely candidate *cis*-acting elements and those patterns that are highly unlikely to be *cis*-acting elements. Furthermore it was predicted that the more likely *cis*-acting element would have a considerably higher score than the non-specific motifs identified, thereby creating a distinct gap in the Total Scores, between candidate and non-candidate motifs. As anticipated this general trend was observed although the effect was less dramatic than originally hoped for (Figure 5.5). The difference in score between the highest scoring motif and the second highest was 6%. The scoring strategy was successful in that the motif ATCGATCA identified with the highest score has high sequence complexity, was over-represented in the positive set and it only occurred in the intergenic region.

The Total Scores of motifs was subsequently modified (to give the Final Score) to incorporate the allowance of degeneracy at a single position within the motif. It was considered important that a motif with a number of variations should be given a higher weighting as they were more likely to represent a *cis*-acting element. The motifs were then ranked based on their Final Score. After factoring in degeneracy at one position the motif ranked highest on Total Score remained at the same position. However the nine motifs with the next highest Total Scores were reduced to a lower ranking (Section 5.4). This provided further confirmation that the motif with the highest Total Score and Final Score was likely to be a *cis*-acting element.

The highest scoring motif ATCGATCA was then used to search in the Transfac Database, which contains all the known *cis*-acting elements (Wingender *et al.*, 1997), however it was not recognized thus indicating it was novel element. In addition searching with this element within ACeDB (local version) demonstrated that it was not part of a repeat element. The expression patterns of the genes containing this motif within the 1 kb upstream region were examined and were found to have no other components of expression that were similar.

The data presented in this chapter suggests that the scoring strategy was successful in identifying a potential candidate *cis*-acting regulatory element. However, as described in the next chapter, it was considered necessary to further evaluate the scoring strategy on a data set in which a previously described *cis*-acting regulatory element had already been verified experimentally.

Chapter 6

Chapter 6: Evaluation of the scoring strategy used to identify candidate *cis*-acting regulatory elements.

6.1: Introduction.

A scoring strategy, based on biological variables, has been developed for the identification of potential *cis*-acting regulatory elements which drive gene expression in the excretory cell of *C. elegans*. As a means to assess the efficacy of this strategy it was considered necessary to test the scoring approach on data sets in which *cis*-acting elements had previously been identified. Two well-characterized regulatory elements initially considered for this purpose were the heat shock element "aGAAtaTTCtaGAAt" (Drees *et al.*, 1997) and the PHA-4 binding site element "TRTTKRY" (R = A/G, K = T/G, Y = T/C; [Gaudet and Mango, 2002]). Unfortunately both these elements have significant disadvantages which limit their use as positive controls. The heat shock element is long at 15 bp, and therefore potentially easier to identify than a shorter motif, however it contains three gaps in between conserved regions which potentially make identification by SPEXS and the scoring strategy extremely difficult. Due to the high sequence variability, the PHA-4 binding site element would occur, by chance, approximately once every 1.6 kb or possibly at an even higher incidence given the 3 Ts at the start of the motif and the AT-rich nature of the intergenic region of the *C. elegans* genome. Therefore the low specificity of this motif may prove to be a problem in its use as a positive control.

More promisingly however Guhatharkurta *et al.* (2002b) recently identified a *cis*-acting regulatory element which directs gene expression in the *C. elegans* muscle cell. The 14 bp muscle specific *cis*-acting element "cccgCGGGagcccg" was identified from a 2 kb region of DNA upstream from the translational start site by a computational approach and was found to contain a highly conserved core region. The authors of the study reported that of the 25 genes that were known to contain the element, 14 genes were found to express in muscle cells. Based on the findings of this study it was considered that the muscle data set would be useful in evaluating the novel scoring strategy.

To affect a valid comparison with the original study by Guhathakurta *et al.* (2002b), the initial SPEXS analysis of the muscle gene set was to be performed using the 2 kb

upstream region of genes in the positive and negative data sets. The positive set consisted of 19 genes (Table 2.7) all of which express in muscle as described by Guhathakurta *et al.* (2002b). The only distinctions in methodology between this and the earlier study being: 1) the negative gene set (Table 2.8) is smaller (90 compared with 3000 genes), and 2) the negative gene set is not randomly selected, but composed of genes known to express in cell types other than the muscle (based on reporter gene fusion analyses previously performed in the Hope laboratory). In addition, as only 2 kb of DNA sequence upstream from the translational start is used for analysis, the scoring system "weightings" of frequency, position and complexity, had to be modified. This is because the specific location of the motif is no longer a critical factor as the analysis is performed on the region of DNA expected to contain the *cis*-acting element. Therefore of the three variable factors the frequency will be given the greater significance. The revised weightings are 70% for frequency and 15% each for position and complexity.

Following the analysis of the muscle gene set, a comparable analysis will then be performed with the excretory cell gene set. By using only the 2 kb DNA from the upstream regions of genes (rather than the entire genomic insert; as used in Chapter 5) it is hoped that specificity will be improved. In addition, as the muscle element was located within the 2 kb upstream regions of the respective genes, it is anticipated that the regulatory elements of excretory cell genes may also be present in these regions. Further support for this hypothesis has been provided by recent promoter deletion and other *cis*-acting element detection studies (Frith *et al.*, 2001; Zhang *et al.*, 2002). (The excretory cell data set used in this analysis will be subsequently referred to as the 2 kb excretory set).

6.2: Defining a cut-off criteria for the number of motifs to be analysed based on frequency distribution for the muscle gene set.

By selecting motifs that occurred on ≥ 6 occasions in the positive set, and also occurred on ≥ 4 more occasions in the positive than in the negative set, only 24 potential muscle-specific motifs were identified (importantly the muscle element being sought was present within this set of motifs). Removing reverse complements subsequently reduced the number of motifs to twelve (emboldened motifs in Appendix III). The low number of motifs satisfying the above criteria was a consequence of a greater number of genes included in the negative set which was in turn due to the use of less computationally-demanding, shorter DNA sequences. To ensure that the scoring strategy was rigorously evaluated and that the muscle element in question could be identified from a large number of potential background motifs, a higher frequency of occurrence in the negative set was tolerated. A 2-dimensional matrix (19 [number of genes in the positive data set] x 90 [number of genes in the negative data set]) was generated to examine the number of different motifs identified by SPEXS at different frequencies in the positive and negative sets (Table 6.1). Selecting the ten (or fewer) motifs occurring the least number of times in the negative gene set for each frequency of occurrence in the positive set generated a convenient number of motifs for further analysis. This resulted in the inclusion of a sufficient number of motifs that were unlikely to be *cis*-acting elements such that the scoring strategy would be thoroughly tested, whilst not allowing too much non-specific noise into the interpretation of the data set (Table 6.1).

Frequency in muscle negative gene set	*Frequency in muscle positive gene set*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
0	70892	27561	506	44	14	**1**	**0**
1	44168	5058	614	90	8	**2**	**0**
2	18060	3799	704	141	38	**2**	**0**
3	10084	2866	734	138	42	8	**2**
4	6278	2267	736	219	48	4	**4**
5	4366	1898	710	204	66	18	**4**
6	3080	1389	618	239	74	22	8
7	2050	1232	602	248	66	34	10

Table 6.1. A section (7x7) of a 2-dimensional matrix (19 x 90) representing the number of motifs generated at different frequencies in the muscle gene sets. The values represent the number of motifs detected at different frequencies in positive and negative sets. The motifs used for further analysis are emboldened and shaded.

6.3: Evaluation of Total Scores.

In total 100 motifs were identified from the SPEXS output which satisfied the cut-off criteria described above. The Frequency, Position and Complexity Values were obtained, for the 100 motifs satisfying the cut-off criteria (as described in the previous Chapter, [Section 5.2.1, 5.2.2 and 5.2.3]). In addition, the formulas for the calculation of the Frequency Score, Position Score and Complexity Score were identical to that used in the genomic insert approach (Sections 5.2.4, 5.2.5 and 5.2.6) although the weighting for each of the biological factors was modified as stated earlier (Section 6.1). As the SPEXS analysis was performed on the region of DNA (2 kb upstream region) predicted to contain the *cis*-acting elements, all motifs had a Position Score of a minimum of 10 %. A typical example of a motif with the lowest Complexity Value (0.45) was GCCCCC and the highest (1.40) was GCAATTGC. As before the Total Scores for each of the motifs were determined by summing the Frequency, Position and Complexity Scores.

6.3.1: Comparison of previously defined muscle elements with those identified using the scoring strategy.

The motifs with the highest Total Scores were TCCCGGGA (93.6) and CGGGATCC (93.5), respectively (Appendix III). Both these motifs were related to the longer muscle motif "cccgCGGGagcccg" identified by Guhathakurta *et al.* (2002b) and contained the central CGGG core. The motifs TCCCGGGA and CGGGATCC occurred in three, and five different genes in the positive set, respectively, with neither motif occurring in the negative set. All occurrences of motif TCCCGGGA were within the 1 kb upstream region. The motif TCCCGGGA was a palindrome of the core region whilst the motif CGGGATCC was similar to the original motif identified by Guhathakurta and colleagues but was shorter and contained a mismatch at a single position as shown below:

```
cccgCGGGagccg   motif identified by Guhathakurta et al. (2002b)
    CGGGatcc    second highest scoring motif.
```

Of the 100 motifs satisfying the cut-off criteria, 19 could be considered part of the original muscle motif (not including the palindromic motif TCCCGGGA), with some motifs containing single base differences (Figure 6.1). Most of the motifs were detected within the 1 kb upstream region. The Complexity Values for all variations of the

muscle motifs averaged 0.83, with the lowest being 0.45 (CCCGCC) and the highest being 1.27 (CGGGATC). Of the 19 motifs considered to be part of the muscle specific *cis*-acting element, three (CGGGATCC, GCGGGAGC and GGGAGCCC) were only detected in the positive set while the other motifs were also detected in the negative set but usually at a lower frequency. However four of these motifs (CCGCGC, GCCCGC, CGGGAG and CCCGCC) occurred with a higher frequency in the negative than in the positive set although they were short in length (6 bp) and therefore more likely to occur by chance than a longer motif.

```
cccGCGGGagcccg Motif 1 as identified by Guhathakurta et al. (2002b)
cccgcggg
cccgcgg
cccgcc
 ccgcgg
 ccgcgc
  cgcgggag
  cgcggga
  cccggga
  ggcggga
   gcgggagc
   gcgggc
    cgggag
    cgggatcc
    cgggatc
    cgggagc
     gggagccc
     gggagcg
     gggcgc
```

Figure 6.1. Sequence alignment of all motifs that could be considered part of the motif identified by Guhathakurta *et al.* (2002b).

In addition to Motif 1 (as named by Guhathakurta and colleagues), four other motifs identified by the original study were also present in the 100 motifs satisfying the cut-off criteria in the current study. These motifs described as Motifs 2, 3, 4 and 5, are CTCTcaaaccc, aAGAAGAagc, TGGGcGGa and ggGCGGGa, respectively were also detected but obtained much lower scores (Appendix III).

6.3.2: The incorporation of degeneracy into the Total Scores for the muscle set.

Originally it was hypothesized that a true *cis*-acting element would possess a significantly greater score than the remainder of the identified motifs which could be considered background noise (non-specific). However, analysis of Total Scores

revealed that this was not the case (Figure 6.2) and there was a relatively gradual incremental change in scores between those ranked highest and lowest. However, after incorporating degeneracy (to give the "Final Score") the score of the top two motifs was much higher than the third and subsequent motifs.

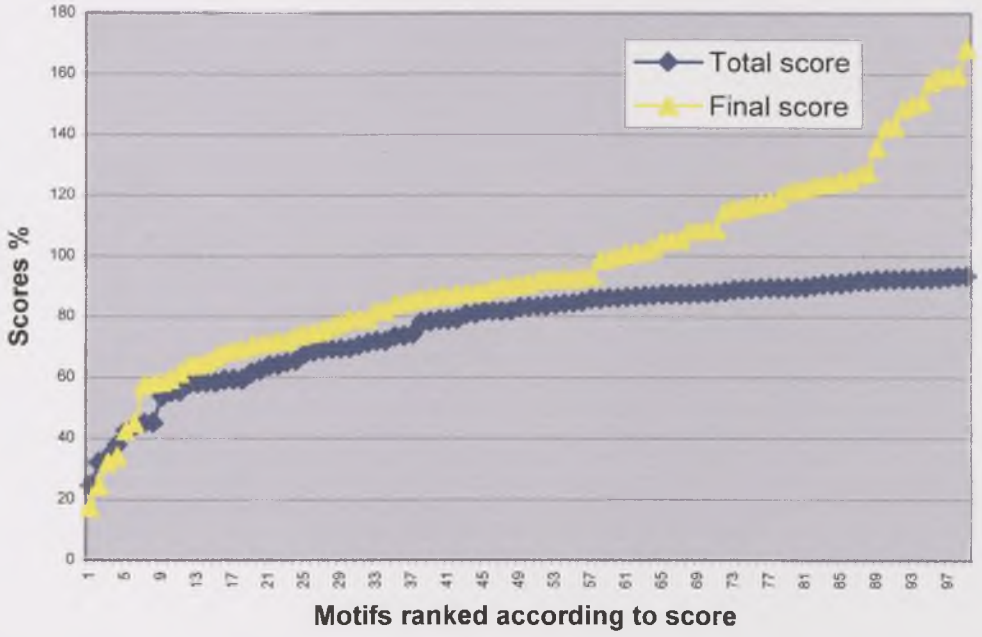

Figure 6.2. Distribution of the Total and Final Scores calculated for motifs obtained from the muscle data set. The plot of the Total Scores demonstrates a gradual increase (thereby making it difficult to distinguish a genuine element from a non-specific motif). However, with the incorporation of degeneracy to give the Final Score, there was a greater difference between the highest scoring motif and remainder of the motifs, thus making it easier to identify an authentic motif from background non-specific noise.

After incorporating degeneracy the motif with the highest Final Score was AGCAATTGCTG. However, inspection of this motif revealed it to be part of a palindrome (CAGCAATTGCTG) and hence the frequency values were artificially increased by almost two-fold as each occurrence of the sequence was counted in both forward and reverse orientations. Adjusting the frequency values for this motif, and other similar motifs which were found to comprise palindromes reduced their Total Scores (as well as the Final Scores incorporating degeneracy). After these adjustments the motif CCCGGGA had the highest Final Score.

Interestingly, incorporating degeneracy (Section 5.4) altered the rankings of the top twenty motifs such that the motifs now ranked second to fifth, ninth, eleventh and twelfth were all variations of the original muscle motif. Thus, incorporating degeneracy rewarded all motifs that were similar, and therefore more likely *cis*-acting elements, with higher scores.

6.4: Analysis of the muscle data set with MEME.

The muscle gene set was subsequently analysed by MEME to determine if this software could detect the previously characterized muscle element. The MEME options used were –mod TCM, -W 12 and -nmotifs 10. The muscle data set used for the MEME analysis was identical to that used in the SPEXS analysis described above. The muscle element was not detected using the TCM option of MEME, however when the experiment was repeated using the ZOOPS option a variation of the original muscle motif (cccgCGGGagcccg) was detected but only ranked six of the ten motifs identified (Table 6.2).

Motifs identified in muscle positive gene set
GGGCGGGGGAAG T G C AAGG A C C
CTTCCCCCGCCC CCTT G C A G G T
TTTTTTCAAAAA CGG
GGAAAAGGGAGG A GGGG G T
GCCCCCCCCCCC TT G T TT A G
CTCCCGCGGGAG **A TA** **T**
CACACGCAAACA G G GA C A G
GTGTGTGTGTCT CG C CC GC
AAAAAAAAGAAA G T
TTCTTTTTTTTC GT C C T

Table 6.2. The ten highest ranking motifs identified by MEME from the muscle gene set using options –W 12, -nmotifs 10 and ZOOPS. The emboldened motif is the muscle *cis*-acting element identified by Guhathakurta *et al.* (2002b).

6.5: Applying the scoring strategy to the excretory data set.

The DNA sequence representing the 2 kb upstream regions of excretory cell-expressing genes was obtained using the protocol described in Section 2.2.2. The negative set (Table 2.5) contained a greater number of genes (n=134) than had been used in the genomic insert approach. This was because analysis of the shorter sequences now being used was less computationally-demanding and therefore more genes could be included. As before the negative set consisted of genes from the Hope Lab Expression Pattern Database, selected on the basis that they expressed in cell groups other than the excretory cell. Genes with ubiquitous expression were not used in the negative gene set as they also express in the excretory cell. Since the earlier analyses of the entire genomic DNA inserts, two additional genes (*T21B10.5* and *Y46G5A.4*) were found to express in the excretory cell and could now be included in the positive set (Table 2.4).

A 2-dimensional matrix of (20 x 134) was generated and a value of nine was selected as the threshold for the number of motifs occurring least frequently in the negative set for each frequency of occurrence in the positive set. All motifs above the threshold were used to screen the positive and negative 2 kb excretory data sets to obtain frequency of occurrence and position data. These data were subsequently used to calculate the Total Scores for each of the motifs in the manner described above (Section 6.2).

6.6: Comparison of motifs detected in the 2 kb excretory set to those obtained from the genomic DNA insert analysis (Chapter 5).

The motif with the highest score from the 2 kb excretory set was TTACCGAA (Table 6.3). It occurred on five occasions in the positive set and once in the negative set. Three of the occurrences of this motif in the positive set were within the 1 kb region (*F54C9.7*, *Y46G5A.4* and *Y62E10A.1*) and the other occurrences were just beyond the 1 kb region (*F10B5.1* and *Y18D10A.23*). This motif was used to screen the list of motifs obtained from the genomic insert approach and although an exact motif was not found several motifs contained the core ACCG region. The two motifs from the genomic insert approach with the greatest number of common bases were CTACCGTAT and TACCGTC, however, they were lowly-ranked, at positions 202 and 240, respectively (Appendix II). The highest-ranking match to TTACCGAA in the genomic insert motifs was ACCGTAAC (ranked 131). As expected the majority of the motifs detected in the 2 kb excretory set were also obtained by the genomic insert approach however the

rankings in the respective data sets varied markedly. These discrepancies in rank were probably due to the high level of non-specific sequences present in the latter data set. It is likely that the entire genomic DNA insert contains too much irrelevant sequence for the scoring strategy to detect a real element. In addition the highest scoring motif (ATCGATCA) identified from the genomic insert approach was subsequently found on nine occasions in the much larger (n=134 as opposed to n=33) 2 kb excretory negative set. It is therefore likely that had it been possible to increase the number of genes in the negative set of the genomic insert approach, the occurrence of this motif would have been higher and as a consequence the motif would ultimately have been assigned a much lower score.

		Positive set		Negative set			Position			Complexity		
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Frequency Score	>1kb	<1kb	Score	Value	Score	Total Score
1	ttaccgaa	5	5	1	1	70	2	3	13	1.321	14.2	97.16
2	agtcgaat	6	7	3	3	69.5	3	4	12.9	1.321	14.2	96.47
3	aacatgttc	4	4	0	0	70.0	2	2	12.5	1.311	14.0	96.47
4	agtatact	3	3	1	1	68.4	0	3	15.0	1.255	12.9	96.30
5	cgaaatttcg	2	2	0	0	68.4	1	1	12.5	1.366	15.0	95.87
6	ccatgttc	6	6	3	3	69.2	4	2	11.7	1.321	14.2	95.01
7	cataacaata	4	4	0	0	70.0	2	2	12.5	1.089	9.8	92.33
8	aaaaagatca	4	4	0	0	70.0	1	3	13.8	0.940	7.1	90.81
9	aatttcagc	9	10	13	14	63.5	4	6	13.0	1.311	14.0	90.43
10	aaacatcat	5	5	2	2	69.2	2	3	13.0	0.995	8.1	90.26
11	taccgaa	8	9	13	13	62.9	3	6	13.3	1.277	13.3	89.59
12	atcgatc	10	11	18	18	60.5	5	6	12.7	1.352	14.7	87.93
13	gaataatta	6	6	4	4	68.4	4	2	11.7	0.937	7.0	87.03
14	aacatcat	9	9	12	12	64.3	4	5	12.8	1.040	8.9	85.97
15	attcgac	10	12	21	21	58.3	6	6	12.5	1.352	14.7	85.52
16	aaggcga	8	8	11	11	64.3	4	4	12.5	1.004	8.3	85.03
17	aaaacatca	7	7	7	7	66.7	3	4	12.9	0.849	5.3	84.94
18	aaagcaaaac	4	4	0	0	70.0	4	0	10.0	0.802	4.5	84.47
19	attgaataa	7	8	8	8	66.2	7	1	10.6	0.937	7.0	83.81
20	agcaaaac	11	12	14	14	64.6	8	4	11.7	0.900	6.3	82.53
21	aataaattaa	5	5	2	2	69.2	3	2	12.0	0.611	0.9	82.09
22	aagatcaa	10	10	18	19	59.9	5	5	12.5	1.074	9.5	81.97
23	aataattaaa	5	5	2	2	69.2	4	1	11.0	0.611	0.9	81.09
24	aattgaata	9	10	14	14	62.9	8	2	11.0	0.937	7.0	80.91
25	aatatatat	9	12	11	11	65.9	6	6	12.5	0.687	2.3	80.75
26	aataaatta	8	8	7	8	67.3	5	3	11.9	0.637	1.4	80.54
27	acattgaa	12	13	26	28	55.0	5	8	13.1	1.213	12.1	80.24
28	aataattaa	7	8	7	7	67.0	6	2	11.3	0.637	1.4	79.64
29	aaaagaaat	8	8	13	13	62.6	2	6	13.8	0.684	2.3	78.67
30	attgaata	12	15	22	22	59.4	11	4	11.3	0.974	7.7	78.40
31	aaattaattt	6	8	12	13	62.1	4	4	12.5	0.693	2.4	77.05
32	acgcac	13	13	32	35	50.4	5	8	13.1	1.011	8.4	71.85
33	atttatca	11	12	30	30	51.5	6	6	12.5	0.974	7.7	71.67
34	gaataca	14	18	39	41	46.8	7	11	13.1	1.154	11.0	70.94
35	aaaacatc	11	13	30	32	51.2	5	8	13.1	0.900	6.3	70.59

Table 6.3. Motifs identified from the 2 kb excretory data set. Frequency of occurrence, position and complexity data (including the scores calculated for each variable) as well as the Total Scores are shown.

		Positive set		Negative set		Frequency Score	Position			Complexity		Total Score
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq		>1kb	<1kb	Score	Value	Score	
36	ctatag	8	9	28	28	43.6	5	4	12.2	1.330	14.3	70.12
37	aatatata	13	18	28	30	55.8	10	8	12.2	0.662	1.9	69.92
38	ataaatta	13	14	28	28	54.2	7	7	12.5	0.662	1.9	68.56
39	taattaaa	11	12	32	35	53.1	6	6	12.5	0.662	1.9	67.47
40	ataattta	12	15	44	51	50.4	4	11	13.7	0.693	2.4	66.50
41	aagatca	17	17	44	50	42.8	9	8	12.4	1.154	11.0	66.16
42	aaatttgg	15	18	33	35	42.2	3	15	14.2	1.082	9.7	66.09
43	aataatta	12	16	44	48	50.1	10	6	11.9	0.662	1.9	63.85
44	atcaatc	15	16	56	62	42.2	12	4	11.3	1.079	9.6	63.11
45	gatcga	18	20	51	60	34.6	7	13	13.3	1.330	14.3	62.16
46	agttcg	15	17	32	39	35.4	9	8	12.4	1.330	14.3	62.08
47	atttaaat	9	11	34	37	46.6	6	5	12.3	0.693	2.4	61.30
48	aatatatt	8	9	46	50	44.9	4	5	12.8	0.693	2.4	60.17
49	aactata	14	15	63	66	39.8	8	7	12.3	0.956	7.3	59.45
50	gacata	17	19	54	62	28.9	8	10	12.8	1.242	12.7	54.34
51	gcaaaac	16	18	51	53	34.0	8	10	12.8	0.956	7.3	54.17
52	aaatatat	14	24	56	61	38.7	12	12	12.5	0.662	1.9	53.03
53	aattttgt	16	20	59	65	33.8	11	9	12.3	0.900	6.3	52.34
54	aaacatc	16	20	56	69	31.1	6	14	13.5	0.956	7.3	51.90
55	agatag	17	18	52	56	31.6	12	6	11.7	1.011	8.4	51.65
56	aagagag	14	18	49	94	35.7	10	8	12.2	0.683	2.3	50.16
57	atatata	20	34	54	58	34.6	16	17	12.6	0.683	2.3	49.42
58	aaaaagag	17	19	74	89	36.0	8	11	12.9	0.562	0.0	48.85
59	atcagc	21	25	67	86	20.4	10	15	13.0	1.330	14.3	47.75
60	aaaaatcg	18	24	74	94	23.2	10	14	12.9	1.074	9.5	45.61
61	agttgc	19	22	67	81	17.2	7	15	13.4	1.330	14.3	44.89
62	atagga	18	21	72	86	23.7	14	7	11.7	1.011	8.4	43.75
63	atttatc	18	21	84	102	19.6	11	10	12.4	0.956	7.3	39.34
64	atcaata	20	26	80	98	11.2	11	15	12.9	0.956	7.3	31.40
65	agcaaaa	19	27	84	102	14.2	14	13	12.4	0.796	4.4	30.94
66	gttaaaa	20	24	84	111	10.6	13	11	12.3	0.956	7.3	30.26
67	atctac	19	22	85	107	7.1	9	13	13.0	1.099	10.0	30.05
68	aaatcat	20	25	98	110	9.0	14	11	12.2	0.956	7.3	28.53
69	tgataaa	19	23	28	28	0.0	10	13	12.8	0.956	7.3	20.17

Table 6.3 continued. Motifs identified from the 2 kb excretory data set. Frequency of occurrence, position and complexity data (including the scores calculated for each variable) as well as the Total Scores are shown.

6.7: Incorporating degeneracy in the analysis of the 2 kb excretory set

By allowing for degeneracy (Section 5.4) at a single position the motif ranked highest by Total Score was reduced to the second highest position and the motif with the third highest Total Score moved up to the top position (Tables 6.4 and 6.5). Other than this change there were no alterations in the ranking of the top ten motifs after the allowance for degeneracy (Table 6.5).

		Positive set		Negative set		Frequency	Position			Complexity		
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Score	>1kb	<1kb	Score	Value	Score	Total Score
1	ttaccgaa	5	5	1	1	70.0	2	3	13.0	1.321	14.2	97.16
2	aacatgttc	4	4	3	3	70.0	2	2	12.5	1.311	14.0	96.47
3	agtcgaat	6	7	0	0	69.5	3	4	12.9	1.321	14.2	96.47
4	agtatact	3	3	1	1	68.4	0	3	15.0	1.255	12.9	96.3
5	cgaaatttcg	2	2	0	0	68.4	1	1	12.5	1.366	15.0	95.87
6	acatgttc	6	6	3	3	69.2	4	2	11.7	1.321	14.2	95.01
7	gataacaata	4	4	0	0	70.0	2	2	12.5	1.089	9.8	92.33
8	aaaaagatca	4	4	0	0	70.0	1	3	13.8	0.94	7.1	90.81
9	aatttcagc	9	10	13	14	63.5	4	6	13.0	1.311	14.0	90.43
10	aaacatcat	5	5	2	2	69.2	2	3	13.0	0.995	8.1	90.26

Table 6.4. **Top ten ranking motifs obtained with the 2 kb excretory set.**

		Positive set		Negative set		Frequency	Position			Complexity			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Score	>1kb	<1kb	Score	Value	Score	Original_rank	Final Score
1	aacatgttc	4	4	3	3	70.0	2	2	12.5	1.311	14.0	2	128.14
2	ttaccgaa	5	5	1	1	70.0	2	3	13.0	1.321	14.2	1	127.02
3	agtcgaat	6	7	7	3	69.5	3	4	12.9	1.321	14.2	3	124.98
4	aataattaa	5	5	2	2	67.3	4	1	11.0	0.611	1.4	28	121.2
5	attgaata	12	15	22	22	72.7	11	4	2.3	0.974	7.7	29	119.58
6	acatgttc	6	6	3	3	69.2	4	2	11.7	1.321	14.2	6	119.13
7	aaacatcat	5	5	2	2	69.2	2	3	13.0	0.995	8.1	10	118.92
8	aataaatta	8	8	7	8	67.3	3	5	11.9	0.637	1.4	26	118.21
9	ataaatta	13	14	28	30	54.2	7	7	12.5	0.662	1.9	38	117.57
10	taccgaa	8	9	13	13	62.9	3	6	13.3	1.277	13.3	11	113.88

Table 6.5. **Top ten ranking motifs obtained with the 2 kb excretory set factoring in degeneracy at one position within the motif.**

6.8: Analysis of the highest scoring motif (TTACCGAA) identified from the 2 kb excretory set.

To determine if any further common components of expression (other than in the excretory cell) was associated with TTACCGAA-containing genes, the 2 kb excretory set was searched to identify all genes possessing the motif. In total 5 genes (*F10B5.1*, *F54C9.7*, *Y62E10A.1*, *Y46G5A.4* and *Y18D10A.23*) contained the motif, and all of the genes with the exception of *Y18D10A.23* being able to drive reporter gene expression in the pharynx and intestine (Table 6.6). The hypodermis was a further component of expression common to four of the genes (*F10B5.1*, *F54C9.7*, *Y18D10A.23* and *Y46G5A.4* but not *Y62E10A.1*). These findings imply that if the motif is a true *cis*-acting element it is likely to be involved in excretory cell gene expression as this cell type is the only component of expression common to all the genes containing the motif.

The motif TTACCGAA was also detected in one gene (*Y47D3A.16*) in the negative set. This gene is expressed in the pharynx and hypodermis as well as the E-lineage, which differentiates into the intestine. The apparent lack of expression of *Y47D3A.16* in the excretory cell may be because expression is weak and is not detectable by the reporter gene fusion approach.

Genes	*Expression pattern descriptions*
F10B5.1	Expression in this multi-component pattern is first seen in late embryogenesis and extends into adulthood. A few nuclei stain in the 3-fold embryos, these are probably **pharyngeal** and/or hypodermal. In L1 larvae we can see expression in the hypodermal nuclei of the head and body, and in the nuclei of the pharyngeal muscles. Staining of the **pharyngeal muscle nuclei** and **hypodermal nuclei** is very strong in some L2 larvae and older worms. The varying numbers of nuclei staining may be due to mosaicism. From L2 larvae through to adulthood, we can see expression in the large **intestinal** nuclei, this sometimes extends to non-localised staining of the **intestine**. The final component is exhibited in late larvae and adults. Staining is seen in the nucleus of the **excretory cell** and in the excretory canals extending anteriorly and posteriorly.
F54C9.1	Expression is seen from embryo to adult in this extensive multicomponent pattern. At the 1.5-fold stage staining is seen in 5 as yet unidentified nuclei, with staining becoming more extensive in the 3-fold embryo. In larvae and adults, staining is seen in the **pharynx (including pharyngeal muscle cell nuclei)**, with many **hypodermal nuclei** in the head, tail and body also staining. Expression is also seen in the **excretory cell** nucleus and canals, in the spermathecae, and in the anal muscles.
Y18D10A.23	Diffuse non-nuclear expression is seen in older larvae and adults in three distinct components, i.e. the **excretory cell** and canals, the posterior **hypodermis**, and the uterine wall (ut2 toroidal cells).
Y46G5A.4	Extensive weak diffuse expression is seen in early and late stage embryos. In larvae and adults expression is still diffuse but more mosaic and is seen in a number of tissues, i.e. **excretory cell**, head ganglia, **intestine**, **pharynx**, and developing gonad.
Y62E10A.1	Mosaic expression begins in elongated embryos and continues through to adulthood. Expression is generally diffuse and not nuclear-localised and is seen in the following tissues: **pharynx**, body-wall muscle, **intestine**, and the **excretory cell**.

Table 6.6. Genes possessing the highest scoring motif from the 2 kb excretory set.

Components of expression in cell groups common to all the genes are emboldened.

Also present in the 2 kb excretory set, ranked 11 was motif TACCGAA, a shorter version of the highest-ranked motif. Although this motif occurred with a high frequency in the positive set it was also detected with a high frequency in the negative set and as a consequence obtained a lower Total Score. However this shorter motif contributed to the Final Score of the longer variant and was a contributing factor in maintaining its high position.

6.9: Comparing the SPEXS analysis of the 2 kb excretory data set with outputs generated by MEME and CONSENSUS software packages.

6.9.1: MEME analysis of the 2 kb excretory data set.

The 2 kb upstream regions from the 20 excretory cell-expressing genes were subsequently analysed by MEME to determine whether similar motifs to those detected by SPEXS and the scoring strategy could be obtained by alternative software packages. Specifying the MEME options –mod TCM, -W 12 and –nmotifs 10, identified motifs

that were mostly AT-rich except for the fifth ranking motif, which was high in GC content (Table 6.7). However, this motif was also contained in several of the genes in the negative set. There was no overlap between the groups of motifs identified by MEME and SPEXS. When the analysis was repeated using the –mod ZOOPS option rather than TCM a largely different set of motifs were obtained (Table 6.7) although as previously observed the top four motifs were AT-rich. However there was a greater number of GC-rich motifs generated with the ZOOPS option than with the TCM option but the relative significance of these motifs in terms of the likelihood of being *cis*-acting elements was difficult to assess as they were also detected in many of the genes in the negative set.

Motifs obtained from positive set using the TCM option	***Motifs obtained from positive set using the ZOOPS option***
TTTTTAAATTTT CC A G	AAATTTTGAAAA TTTA CGC AA
TTTTTCATTTTT CCCCATA C G	TCCGATTTTTCC GT A C CC A
TTTTTTTAAAAA CATTTTT G	TTTTTTTTTGAA A G C
GAAAAAGAAAAA AG GG AGGGGG C	TTCAAAAAAAAA GA
GCTCTCGCCGCC CGGGG CAGAGG TC T C	GAAAGAAAGAGA GGGAGG A
TTTCAAAAAAAA CTCC C A G	GGCCTCTCCGCC TCG CG TG CA AT
TTTTTAAAATTT CT T C	GGTGAAGAGAGC C C GC CG G
AAAAAAAAAAAA GG	TCTCTGTTTTTC TCC C CC T A
AAAAAAAAAAAA TCC	GTTTTCGACACG C C TCCAT A C
TTTTTAAATTTT CC A G	GGAAAAATTGAA A G GCC G T

Table 6.7. Motifs identified by MEME from the 2 kb upstream regions of 20 genes which expressed in the excretory cell.

6.9.2: CONSENSUS software analysis of the 2 kb excretory data set.

The original analysis of the promoter regions of the muscle data set by Guhathakurta *et al.* (2002b) was performed with the software CONSENSUS (Hertz and Stormo, 1999) and ANN-Spec (Workman and Stormo, 2000). The CONSENSUS algorithm searches for a matrix with either a low probability of occurring by chance, or, that has a high information content. The ANN-Spec algorithm searches for a weight matrix that maximizes the specificity of binding to a positive sequence set compared to a negative sequence set, which in the case of the study by Guhathakurta and colleagues consisted of 3000 sequences. Both these algorithms generated matrices which were then used to search the muscle data set for all sites containing the motif using an algorithm known as PatSer (http://ural.wustl.edu/~jhc1/consensus/html/Html/main.html). The PatSer program allows the scoring of the "words" of the sequence against a weight matrix.

To determine whether the approach by Guhathakurta *et al.* (2002b) could identify the motifs detected by the current strategy, it was decided to re-analyze the 2 kb excretory set with the software of the original study. Unfortunately the ANN-Spec algorithm could not be used over the World Wide Web (WWW) as the matrices required as prerequisite for genome analyses were only available for the Yeast and the Human genomes and in addition the comparison would not be effective due to the low number of sequences in the negative set. However the CONSENSUS (version 6.c) program was available via the WWW and was performed applying a motif length of 8 bp for both strands of the DNA. The results of this analysis contained eight motif-matrices. However, the first four matrices were identical, and the next four matrices were reverse complements of the first four motif-matrices (Figure 6.3). The consensus motif represented by these eight matrices was not present in the SPEXS output.

```
MATRIX 1
number of sequences = 3
unadjusted information = 13.8131
sample size adjusted information = 9.07699
ln(p-value) = -35.0998   p-value = 5.70625E-16
ln(expected frequency) = -5.27141   expected frequency = 0.00513637
A |   0   0   0   3   0   0   0   0
C |   3   0   0   0   0   0   0   0
G |   0   3   3   0   3   3   3   3
T |   0   0   0   0   0   0   0   0
  1|2   :   1/1394  CGGAGGGG
  2|3   :  16/1615  CGGAGGGG
  3|1   :  20/1266  CGGAGGGG
```

Figure 6.3. The highest ranking motif from the CONSENSUS algorithm analysis of the 2 kb excretory data set. The output generated shows that the number of sequences containing the motif, the information content (the higher the occurrence of the motif in data set the higher the information content value), p-value (statistical measure of similarity, the higher the value the better the similarity between sequences) and the expected frequency of occurrence of the motif by chance (the lower the value the less likely it is to occur by chance). Next, the motif matrix showing the frequency of each base at each position. Lastly, the first occurrence of the motif is in sequence 1 at position 1394 (shown as 1/1394). The numbers, "1|2", indicate the motif is ranked second. The ranking is based on the number of bases the motif deviates from the consensus sequence and the position in the sequence, therefore the closer to the translational start site the higher the ranking.

6.10: Discussion.

6.10.1: Analysis of the muscle data set.

To validate the usefulness of SPEXS and the scoring strategy to identify *cis*-acting elements an analysis was performed with a muscle gene set which was known to contain a previously defined *cis*-acting regulatory element (Guhathakurta *et al.*, 2002b). Subjecting the motifs selected from the SPEXS output to the novel scoring strategy resulted in the identification of multiple variations of the muscle "Motif 1" (cccgCGGGagcccg) determined by Guhathakurta *et al.* (2002b). One of these variations was the highest scoring motif and the majority of the variations of the motif were detected within the 1 kb upstream region. The successful identification of the muscle regulatory element by the application of the scoring strategy to the SPEXS output validated the approach as a means to identify *cis*-acting elements.

Ideally it would have been preferable to test the scoring system against several data sets in addition to the muscle set. However this was not possible due to the lack of previously determined *cis*-acting elements for specific *C. elegans* cell types. Although *cis*-acting elements from co-regulated genes of other organisms have been identified and verified via biochemical characterization, they were not of use in this analysis since it was considered likely that different properties of the genomes would make such a test irrelevant for the situation in *C. elegans*. It is likely that the weightings of the three biological factors incorporated into the scoring strategy would have to be optimized for a particular species.

The muscle *cis*-acting element was also detected by the MEME software using the ZOOPS option however it was ranked six out of ten. The core of the muscle element is well conserved and therefore the element should be detected easily.

6.10.2: Analysis of the 2 kb excretory data set.

As the scoring strategy had permitted the successful identification of the muscle *cis*-acting element, a similar analysis was repeated with the *C. elegans* excretory set, however using the 2 kb upstream regions of the respective genes (rather the entire genomic DNA inserts as had been used in the reporter gene fusion approach; Chapter 5). In addition because the sequences were now much shorter it was possible to include a greater number of genes in the negative set than had been used in the genomic insert

approach (Chapter 5). A large negative set was considered advantageous to increase the likely specificity of those motifs over-represented in the positive set.

After applying the scoring strategy, the motif with the highest Total Score was TTACCGAA. It occurred with a frequency of five (*F10B5.1*, *F54C9.1*, *Y18D10A.23*, *Y46G5A.4* and *Y62E10A.1)* and was found within the 1 kb upstream region of three of the genes (*F10B5.1*, *F54C9.1* and *Y62E10A.1*). All five genes had multi-component expression patterns and 4 of the 5 genes had additional expression in the pharynx, intestine and hypodermis. This motif was detected in only one of the negative set genes (*Y47D3A.1.6*) and it is possible that this gene does express in the excretory cell but at a low level such that it is not detected using the reporter gene fusion approach.

It was anticipated that the potential *cis*-acting motif (TTACCGAA) would occur with a higher frequency in the positive gene set than was observed (i.e. only 5 of the 20 genes). A possible explanation of the low occurrence is that SPEXS only identifies motifs that are exact matches. Therefore there may be several variations of a motif that occur with a low incidence and as a consequence may not meet the required threshold criteria and are thus absent from the list of the motifs in the SPEX output.

Factoring in degeneracy at a single position into the scoring system lowered the highest-ranking motif to the second highest. However, the difference between the Final Scores of the top two motifs was very small. In addition there was a further motif (TACCGAA) present in the data set that was shorter by 1 bp but otherwise identical to the highest scoring motif TTACCGAA. The presence of the two motifs with very similar sequences contributed to the Final Score of both motifs. As the highest-ranking motifs from this analysis were over-represented in the positive set and were present at locations where elements are predicted to occur, they were considered good candidates for *cis*-acting elements.

As the application of the scoring system to the SPEXS output resulted in the successful identification of a previously characterised *cis*-acting element, this suggests that the motif TTACCGAA, identified from the promoters of excretory cell-expressing genes by the same approach may also be a *bona fide* regulatory element. Therefore in the next

chapter of this thesis the presence of this motif will be analysed in additional data sets to further evaluate its role as a potential *cis*-acting element.

Chapter 7

Chapter 7: *In silico* testing of candidate *cis*-acting elements and the identification of novel genes predicted to express in the excretory cell.

7.1: Introduction.

To provide further *in silico* evidence of a possible role as a *cis*-acting element, each of the candidate motifs identified (by SPEXS and the scoring strategy) from the 2 kb excretory set (Section 6.6) and the genomic insert approach (Chapter 5) were further evaluated by analysis of;

1) a data set of excretory cell-expressing genes not used in any previous analyses, described as the "excretory test set".

2) a data set of *Caenorhabditis briggsae* orthologues of *C. elegans* excretory cell-expressing genes, described as the "*C. briggsae* test set".

The most likely candidate element will then be used to search within the *C. elegans* genome to identify all genes possessing the particular element within their 1 kb upstream regions.

7.2: The excretory cell test set.

The excretory test set (Table 2.9) consists of 23 genes that express in the excretory cell in addition to other cell types, and includes two genes (*clh-4*and *exc-5*) that express solely in the excretory cell. The excretory test set has not previously been used in this project and contrasts with the 2 kb excretory data set (Chapter 6) in that it contains genes from the published literature and WormBase database rather than being obtained from the Hope laboratory research (as is the case for the latter). The genes in the excretory test set were generated using a number of different experimental techniques; immunostaining, *in situ* hybridization and reporter gene fusion (but not from Hope laboratory). For genes characterized using the reporter gene fusion technique, (obtained from WormBase) the coordinates required for obtaining the genomic DNA insert sequences were not always provided and therefore could not be used in the earlier genomic insert analysis. As a consequence these genes were set aside to be used as a test data set for testing the motifs detected from the 2 kb excretory analyses (Chapter 6). It is anticipated that if the motif is a genuine regulatory element then it should occur with a similar frequency and at similar positions in the excretory test set.

7.3: Comparison of motifs identified from the genomic DNA insert approach and the excretory test set.

The highest-ranking motif from the genomic insert approach was ATCGATCA. However, it was only present at a single site (*F55C7.7/unc-73*) in the excretory test set, suggesting that it is unlikely to direct expression in the excretory cell, even though, it was within the 1 kb upstream region. The gene *unc-73* has a complex expression pattern with expression observed in many cell types including the excretory cell, neurons, pharynx, intestine and the hypodermis. There was another variation of the motif, ATCGATCT, within the list of motifs from the genomic insert approach; however, this variation of the motif was not detected in the excretory test set. The excretory test set analysis provides no support for the results obtained with the genomic insert data set.

7.4: Analysis of results from the 2 kb excretory set.

Two kb DNA sequences were extracted from the upstream regions of all genes in the excretory test set (Section 2.2.2). This data set was then used to obtain frequency of occurrence and the position of each of the motifs identified within the promoters of the genes in the 2 kb excretory set (Chapter 6).

The motif, TTACCGAA attaining the highest score from the 2 kb excretory set was detected in 4 of the 23 genes. Two of the genes (*F33D4.2A/itr-1* and *ZK455.7/pgp-3*) contained the motif within the 1 kb upstream region, whereas *K03D10.1/kal-1* and *R09A10.2/gsa-1* contained the element beyond this region. The motif TTACCGAA was located at position 345 bp upstream from the translational start of *itr-1*. This finding is consistent with the length of the promoter (438 bp of the upstream sequence) (Gower *et al.*, 2001) used in the reporter gene fusion that provided expression in the excretory cell (Table 7.1, experiment 2 of *itr-1* reporter expression pattern). A shorter version of the motif (TACCGAA) was detected in an additional gene, *Y38F2AL.3/vha-11*. Analysis of the expression patterns of these genes revealed that expression in the excretory cell was the only component of expression common to all these genes (Table 7.1).

In total seven of the 69 motifs were not detected in the excretory test set (Table 7.2). Of these seven motifs, four were ranked within the top 10 motifs identified from the 2 kb excretory set. These motifs, AGTCGAAT, CGAAATTTCG, ACATGTTC and GATAACAATA, were ranked second, fifth, sixth and seventh in the 2 kb excretory set, respectively (Table 6.3). The absence of these motifs in the excretory test set suggests that they are non-specific and do not represent *cis*-acting elements regulating gene expression in the excretory cell.

Gene	*Experimental technique*	*Expression pattern descriptions*
R06A10.2/gsa-1	Reporter gene fusion	Extensive expression observed in embryos. Larval stages and adults expression restricted to neural and muscle cells, virtually all neurons stain (head ganglia, ventral nerve cord and tail ganglia). Also hermaphrodite-specific neurons (HSNs) and canal-associated neurons (CANs) show expression. Most muscle cells show expression. Body wall muscle cells show a punctate expression pattern of the translational gsa-1::gfp fusion, which may represent localization in dense bodies. Other muscle cells showing expression include those of the pharynx and uterine and vulva muscle cells. Transcriptional gsa-1::gfp fusion was expressed in the bilateral processes of the **excretory cell** and at low levels in the intestine.
F33D4.2A/itr-1	Reporter gene fusions	1) Expressed in intestine, pharynx, pharyngeal terminal bulb, pharyngeal isthmus, pharyngeal intestinal valve, mu_sph, and **excretory cell** at unspecified stage. Not observed in the nervous system or germ line, which differs from immunostaining results. 2) Expression from pB is observed in the spermatheca, PDA neuron, amphid socket cells and **excretory cell**.
	Immunostaining	Expressed in intestine, pharynx, pharyngeal intestinal valve, mu_sph, and **excretory cell** at unspecified stage(s). Expression in ventral nerve cord, spermathecae and germ line observed with immunostaining but not GFP reporter gene. Pharyngeal results differ slightly from GFP reporter gene results.
K03D10.1/kal-1	Reporter gene fusion	Expression of the reporter is first detected in embryos at around the 50-cell stage in 23 cells and widens during embryonic development. At the comma stage, expression is seen in a set of cells whose position is consistent with neuroblasts in the tail as well as in the head where they form a ring-like structure. In larval stages and throughout adult stages, expression is largely restricted to a set of neurons in several head ganglia (AIY, AIZ, RID, M5, ASI, and subsets of labial sensory neurons), motor neurons in the ventral nerve cord, neurons in the midbody region (HSN, CAN, and PVM) and in the tail ganglia (DVB, DVC, and PDB). Consistent nonneuronal expression could be observed in the **excretory cell** and uterine cells.
ZK455.7/pgp-3	Immunostaining	*pgp-3* is mainly expressed in the apical membrane of the large H-shaped **excretory cell**. Staining was also found in the apical membrane of intestinal cells and in the membrane of the most anterior region of the pharynx.
Y38F2AL.3/vha-11	Immunostaining	In the adult stage, VHA-11 was expressed mainly in an H-shaped **excretory cell** and also in intestinal cells. Diffuse staining in the cytoplasm could be seen at all embryonic stages. Beginning at the comma stage, dense dot-like staining became clearly visible in intestinal cells and was detectable during embryonic development. These results indicate that V-ATPase is localized in intracellular compartments of embryos, especially those of intestinal cells.

Table 7.1. Expression patterns of genes from the excretory test set containing the motif TTACCGA(A) in the 2 kb upstream region. For the gene *itr-1*, two different expression pattern descriptions were obtained (from two different experiments) using the reporter gene fusion method, indicated by the numbers 1 (reporter gene fusion contains 3.2 kb upstream region) and 2 (reporter gene fusion contains 478 bp of upstream region). Expression in cell groups common to all genes is emboldened.

#	Motif	Positive set		Position	
		Frag_Freq	Tot_Freq	>1kb	<1kb
1	ttaccgaa	4	4	2	2
2	aacatgttc	0	0	0	0
3	agtcgaat	0	0	0	0
4	agtatact	1	1	1	0
5	cgaaatttcg	0	0	0	0
6	acatgttc	0	0	0	0
7	gataacaata	0	0	0	0
8	aaaaagatca	1	1	1	0
9	aatttcagc	2	2	1	1
10	aaacatcat	4	4	3	1
11	taccgaa	6	6	4	2
12	atcgatc	2	2	1	1
13	gaataatta	1	1	1	0
14	aacatcat	5	5	4	1
15	attcgac	5	5	4	1
16	aaggcga	5	5	3	2
17	aaaacatca	2	2	1	1
18	aaagcaaaac	0	0	0	0
19	attgaataa	0	0	0	0
20	agcaaaac	2	2	1	1
21	aataaattaa	2	2	1	1
22	aagatcaa	3	3	2	1
23	aataattaaa	2	2	1	1
24	aattgaata	1	1	1	0
25	aatatatat	4	4	3	1
26	aataaatta	3	3	1	2
27	acattgaa	3	3	1	2
28	aataattaa	3	3	2	1
29	aaaagaaat	1	2	1	1
30	attgaata	2	3	1	2
31	aaattaattt	2	2	0	2
32	acgcac	4	4	2	2
33	atttatca	3	3	1	2
34	gaataca	10	11	7	4
35	aaaacatc	3	3	2	1
36	ctatag	5	6	3	3
37	aatatata	8	8	6	2
38	ataaatta	5	5	2	3
39	taattaaa	9	9	3	6
40	ataattta	5	6	3	3
41	aagatca	7	7	4	3
42	aaatttgg	13	17	8	9
43	aataatta	6	7	4	3
44	atcaatc	10	14	8	6
45	gatcga	9	12	5	7
46	agttcg	14	20	10	10
47	atttaaat	2	2	0	2
48	aatatatt	8	10	4	6
49	aactata	16	20	7	13
50	gacata	11	14	6	8
51	gcaaaac	7	11	9	2
52	aaatatat	12	15	7	8
53	aattttgt	6	6	3	3
54	aaacatc	12	17	9	8
55	agatag	12	18	7	11
56	aagagag	13	15	7	8
57	atatata	14	20	8	12
58	aaaaagag	10	10	2	8
59	atcagc	15	21	14	7
60	aaaaatcg	5	12	0	12
61	agttgc	14	16	4	12
62	atagga	15	21	10	11
63	atttatc	7	7	3	4
64	atcaata	10	11	5	6
65	agcaaaa	18	20	10	10
66	gttaaaa	21	27	13	14
67	atctac	18	28	13	15
68	aaatcat	19	25	14	11
69	tgataaa	16	20	14	6

Table 7.2. Motifs used in the excretory test analysis. Frequency of occurrence and position data is shown. Motifs are ranked based on Total Score obtained from the 2 kb excretory set.

7.5: Identification of *C. briggsae* orthologues of *C. elegans* excretory cell-expressing genes.

Selective pressure applied to an organism results in functional elements (such as *cis*-acting elements) evolving at a slower rate than non-functional elements (Blanchette and Tompa, 2002). Therefore functional elements may be recognized by their conservation between two related species (Makalowski and Boguski, 1998) and a comparison of the non-coding regions of homologous genes may reveal conserved regulatory elements (Gilleard *et al.*, 1997, Zhang and Emmons, 2000; Web *et al.*, 2002). This concept has been the premise of comparative analyses between the mouse and human genomes, which has lead to the discovery of the regulatory elements for many genes (Hardison *et al.*, 1997; Jareborg *et al.*, 1999). *C. briggsae* is the most closely related nematode to *C. elegans* (Blaxter, 1998) and it is believed that the species diverged approximately 25-50 million years ago (Ayala *et al.*, 1996). This period of time is generally considered to be of sufficient length for divergence of non-functional elements to occur (Kent and Zahler, 2000; Web *et al.*, 2002). Therefore if a candidate *C. elegans cis*-acting element is well-conserved and detected in the *C. briggsae* data set it provides further evidence that the motif is a genuine element.

The *C. briggsae* test set consists of *C. briggsae* orthologues of *C. elegans* genes from both the 2 kb excretory set (used in Chapter 6 and Table 2.4) and the excretory test set (Section 7.2 and Table 2.9). Orthologues were identified by using *C. elegans* sequence to search the *C. briggsae* genome, with the BLASTX algorithm (Altschul *et al.*, 1990). To confirm orthology the identified *C. briggsae* gene was used to "re-identify" the closest homologue in the *C. elegans* genome. Twelve of the 20 genes in the 2 kb excretory set, and 14 of the 23 genes in the excretory test set, were found to have orthologues in the *C. briggsae* genome. Therefore, in total the *C. briggsae* data set consisted of 26 genes (Table 7.3). A further 14 orthologues were identified, however the methionine start codons could not be aligned thus making it difficult to determine the translational start of the gene in *C. briggsae*. As a consequence these genes were not used in the analysis. Only three genes (*B0285.6*, *C46C2.1* and *R12G8.2*) did not have *C. briggsae* orthologues.

C. elegans genes with potential C. briggsae orthologues	
In the 2 kb excretory set.	***In the excretory test set.***
C17H12.14	*K04G2.8/apr-1*
F10B5.1	*E04F6.11/clh-3*
F41E7.1	*Y75B12B.5/cyp-3*
F54C9.1	*T08G11.2/egl-32*
R05G6.6	*R06A10.2/gsa-1*
R13H4.5	*F33D4.2A/itr-1*
T14E8.1	*K03D10.1/kal-1*
Y18D10A.23	*R107.8/lin-12*
Y46G5A.4	*C09B8.7/pak-1*
Y62E7A.1	*ZK455.7/pgp-3*
Y70G10A.3	*R31.1/sma-7*
Y113G7.24	*R10E11.2/vha-2*
	F35H10.4/vha-5
	Y38F2AL.3/vha-11

Table 7.3. **A list of *C. elegans* genes for which *C. briggsae* orthologues were identified and made up the "*C. briggsae* test set".**

7.6: Analysis of the *C. briggsae* test set.

The 2 kb DNA sequences were extracted from the upstream regions of the genes in the *C. briggsae* test set (Section 2.2.2). The highest ranking motif, TTACCGAA, from the 2 kb excretory set (Chapter 6) was only present in one sequence in the *C. briggsae* test set. However, it was located within the 1 kb upstream region of the *C. briggsae* orthologue of the *C. elegans* gene *clh-4*, (Table 7.4). The sole variant of this motif that was found was TACCGAA and this was detected beyond the 1 kb upstream region in the gene *K04G2.8a/apr-1* (expression pattern in Table 7.6).

Gene	***Expression pattern description***
T06F4.2/clh-4	1) Excretory cell 2) A promoter element from CeCLC-4 directed GFP expression only to a single cell, the large, H-shaped, excretory cell. Expression patterns were identical in all four larval stages and adults except the vulval muscles and the HSN neurons.

Table 7.4. **Expression pattern of the *C. elegans* gene *clh-4*.** The *C. briggsae* orthologue of this gene contains the element TTACCGAA within the 2 kb upstream region. The two expression pattern descriptions indicated by numbers 1 (includes 4 kb upstream region in reporter gene fusion) and 2 (6.9 kb upstream region in reporter gene fusion) were obtained from two experiments using the reporter gene fusion experimental approach.

The second highest scoring motif (AGTCGAAT) identified from the 2 kb excretory set also occurred in only one sequence in the *C. briggsae* set (Table 7.5). Furthermore, approximately 50% of the motifs that were present in the C. *briggsae* test set occurred with a Total Frequency (includes multiple occurrences within a sequence) of four or less (Table 7.5). In total there were ten motifs identified from the 2 kb excretory set that were not detected in the *C. briggsae* test set (Table 7.5). This could be explained by the elements recognized by the excretory cell specific regulatory system being poorly conserved between the genomes of the two species.

#	Motif	Positive set		Position	
		Frag_Freq	Tot_Freq	>1kb	<1kb
1	ttaccgaa	1	1	0	1
2	aacatgttc	0	0	0	0
3	agtcgaat	1	1	0	1
4	agtatact	0	0	0	0
5	cgaaatttcg	0	0	0	0
6	acatgttc	0	0	0	0
7	gataacaata	0	0	0	0
8	aaaaagatca	1	1	0	1
9	aatttcagc	3	3	2	1
10	aaacatcat	1	1	1	0
11	taccgaa	2	2	1	1
12	atcgatc	5	5	4	1
13	gaataatta	1	1	1	0
14	aacatcat	3	4	2	2
15	attcgac	6	6	4	2
16	aaggcga	2	4	0	4
17	aaaacatca	0	0	0	0
18	aaagcaaaac	1	1	1	0
19	attgaataa	2	2	1	1
20	agcaaaac	2	2	1	1
21	aataaattaa	0	0	0	0
22	aagatcaa	4	5	1	4
23	aataattaaa	0	0	0	0
24	aattgaata	3	3	1	2
25	aatatatat	3	3	3	0
26	aataaatta	3	3	1	2
27	acattgaa	4	4	3	1
28	aataattaa	2	2	2	0
29	aaaagaaat	3	3	3	0
30	attgaata	4	4	1	3
31	aaattaattt	0	0	0	0
32	acgcac	6	6	2	4
33	atttatca	7	7	4	3
34	gaataca	8	9	7	2
35	aaaacatc	5	5	2	3
36	ctatag	7	10	5	5
37	aatatata	2	2	1	1
38	ataaatta	5	5	1	4
39	taattaaa	1	1	0	1
40	ataattta	7	7	1	6
41	aagatca	11	13	5	8
42	aaatttgg	6	12	7	5
43	aataatta	2	2	2	0
44	atcaatc	11	14	7	7
45	gatcga	17	17	11	6
46	agttcg	16	20	10	10
47	atttaaat	1	1	0	1
48	aatatatt	0	0	0	0
49	aactata	7	7	5	2
50	gacata	12	14	8	6
51	gcaaaac	5	7	4	3
52	aaatatat	6	6	4	2
53	aattttgt	10	12	5	7
54	aaacatc	10	10	3	7
55	agatag	17	19	3	16
56	aagagag	15	17	5	12
57	atatata	6	6	1	5
58	aaaaagag	13	13	5	8
59	atcagc	17	25	15	10
60	aaaaatcg	12	25	15	10
61	agttgc	14	15	7	8
62	atagga	17	24	12	12
63	atttatc	14	15	9	6
64	atcaata	13	14	8	6
65	agcaaaa	16	21	7	14
66	gttaaaa	14	17	11	6
67	atctac	14	17	8	9
68	aaatcat	18	22	15	7
69	tgataaa	22	25	8	6

Table 7.5. Motifs used in the analysis of the *C. briggsae* test set. Frequency of occurrence and position data is shown. Motifs are ranked based on Total Score obtained from the 2 kb excretory set.

7.7: The screening of the *C. briggsae* test set with motifs identified from the genomic insert approach.

The highest scoring motif from the genomic insert approach, ATCGATCA was detected beyond the 1 kb upstream regions of the *C. briggsae* orthologues of two *C. elegans* genes (*K04G2.8a/apr-1* and *T08G11.2/egl-32*). Both these genes had multi-component expression patterns and included neuronal expression in addition to that in the excretory cell (Table 7.6). The other variation of this motif (ATCGATCT) was not detected. The third highest ranking motif, AAGATATC obtained from the genomic insert approach occurred in six sequences (*C09B8.7/pak-1*, *E04F6.11a/clh-4*, *F33D4.2a/itr-1*, *R06A10.2/gsa-1*, *Y62E10A.1* and *ZK455.7/pgp-3*) with two occurrences in gene *R06A10.2*. However this motif was always located beyond the 1 kb upstream region and it was not detected in the excretory test set. Therefore it was considered unlikely to be involved in mediating expression in the excretory cell.

Gene	*Expression pattern description*
K04G2.8a/apr-1	In early L1 larvae, APR-1 was expressed in the 12 Pn cells as they were descending toward the ventral midline. In L3 stage hermaphrodites, APR-1 was strongly expressed in P38.p, but no APR-1 could be detected in anterior (P1.p and P2.p) or posterior (P9.pP12.p) Pn.p cell. In addition, APR-1 was expressed in the seam cells, in the excretory cell and in the excretory canal cell during the L1 stage, and in some unidentified neurons in the head region. Similar to the staining observed in embryos, some APR-1 staining was localized at the adherens junctions, whereas another fraction of APR-1 could be detected in the cytoplasm.
T08G11.2/egl-32	Intense fluorescence in the nuclei of many somatic cells including neurons, body wall muscle and vulval muscle cells, hypodermal and gut cells and the excretory cell.

Table 7.6. Expression pattern of the *C. elegans* genes *apr-1* and *egl-32*. The *C. briggsae* orthologue of these genes contain the element ATCGATCA within the 2 kb upstream region.

7.8: Comparison of motifs detected in the excretory and *C. briggsae* test sets, the 2 kb excretory set and the genomic insert approach.

The highest-ranking motif from the 2 kb excretory set, TTACCGAA, was detected in all three data sets. It was present in 4/23 genes from the excretory test set but was detected just once in the *C. briggsae* test set (Table 7.7). A shorter version of this motif was detected in an additional gene in both the excretory and *C. briggsae* test sets. The presence of this motif in the excretory test set provides additional support that this motif is of importance in controlling expression in the excretory cell. Examination of expression pattern data of all genes containing the element TTACCGAA from the three different data sets, revealed that excretory cell expression was the only common component of expression to all genes.

The highest-ranking motif ATCGATCA from the genomic insert approach was also detected in all three other data sets however it was detected only once in the excretory test set and twice in the *C. briggsae* set (Table 7.7). The low frequency of occurrence of this motif in the positive test sets suggested that it was unlikely to be a *cis*-acting element. In addition, this motif occurred with a frequency of nine in the larger negative set used in the 2 kb excretory set (Chapter 6) and only other variation of this motif was not present in either of the two test sets. Therefore it was possible to conclude that the motif ATCGATCA is unlikely to be a genuine element.

Genes containing the motif TTACCGAA		
2 kb excretory data set	*Excretory test data set*	*C. briggsae data set*
F10B5.1 *F54C9.7* *Y18d10A.23* *Y46G5A.4* *Y62E10A.1*	*F33D4.2a/itr-1* *R06A10.2/gsa-1* *K03D10.1/kal-1* *Zk455.7/pgp-3*	*E04F6.11/clh-3*
Genes containing the motif ATCGATCA		
Motifs detected within the 2 kb upstream region from genomic insert approach	*Excretory test data set*	*C. briggsae data set*
b0285. 6 c46c2.1 *f44a2.5* *r12c12.6* *r13h4.5* *Y18d10A.23* *Y62E10A.1*	*F55C7.7/unc-73*	*K04G2.8a/apr-1* *T08G11.2/egl-3*

Table 7.7. Genes containing the highest scoring motifs from the 2 kb excretory set and the genomic insert approach in the different data sets.

7.9: The prediction of novel genes which potentially express in the excretory cell.

Searching the six chromosomes of the *C. elegans* genome for the presence of the motif, TTACCGAA (in both orientations) resulted in the detection of 3771 occurrences (Table 7.8), indicating that the motif occurs approximately once every 27 kb of DNA (calculation shown below).

Total number of occurrences of TTACCGAA within the genome = 3771,
Length of genome = 100258171 bases.
Therefore the motif occurs once every (100258171/3771 =) 26586 bases
i.e. approximately once every 27 kb

Assuming an equal proportion of all four bases, the motif TTACCGAA is predicted to occur randomly every 65 kb. However as the *C. elegans* genome (particularly the intergenic region) is AT-rich, the motif is expected to occur once every ~51 kb (calculation shown below).

The *C. elegans* genome is composed of (http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/DNA.shtml): A 32%, T 32%, G 18% and C 18%

The probability the motif TTACCGAA will occur in the *C. elegans* genome = $0.32x0.32x0.32x0.18x0.18x0.18x0.32x0.32 = 1.95x10^{-5}$

As a consequence the chance occurrence of the motif is 1/51 kb.

As demonstrated above the motif occurs on average once every 27 kb, although based on sequence composition of the genome this motif would be expected to occur only once every 51 kb. Furthermore in the positive 2 kb excretory data set the motif occurred on 5 occasions in a total of 40 kb of DNA (i.e. on average once every 8 kb), it is therefore possible to conclude that the motif in question is highly enriched in the positive set when compared to the remainder of the genome.

In total 447 genes contained the motif, TTACCGAA, within the 1 kb upstream region. Of these genes five had previously been used in the SPEXS analysis (2 kb excretory set and excretory test set) (Table 7.9). Of the remaining 442 genes, only four genes had fully-characterized expression patterns although none of these genes expressed in the excretory cell.

The remaining 438 genes that contained the motif were novel and no gene expression pattern information was available. Analysis of gene ontology data suggested that the majority of these genes have a role in cation, hydrogen and sodium transport, metabolism, ATP binding or are involved in phosphorylation. The putative biological roles of these genes are consistent with the function of the excretory cell in regulating osmotic and ionic balance (Wang and Chamberlin, 2002; Bürglin and Ruvkun, 2001).

Chromosome	*Frequency of occurrence*
	TTACCGAA
I	657
II	561
III	570
IV	621
V	782
X	580
Total	**3771**

Table 7.8. Total frequency of occurrence of the motif TTACCGAA detected in the six chromosomes of the *C. elegans* genome.

Genes containing the motif TTACCGAA/TTCGGTAA within the 1 kb upstream regions												
	Chromosome I (+)	*I (-)*	*Chromosome II (+)*	*II (-)*	*chromosome IIII (+)*	*IIII (-)*	*chromosome IV (+)*	*IV (-)*	*Chromosome V (+)*	*V (-)*	*chromosome X (+)*	*X (-)*
1	C09D4.3	B0041.2a	AH6.8	C01F1.2	B0464.3	3R5.2	C08F11.3	B0001.1	C01B4.3	B0238.13	C02F12.9	C05E11.6
2	C09D4.6	C01G8.6	C28F5.5	C08G5.3	F11H8.4b	C07H6.3	C27D8.2	C05C12.4	C02A12.6	B0250.4	C11G6.4	C10E2.4
3	C15C6.1	C43E11.5	C40A11.8	C13A10.1	F17C8.3	C14B1.9	C27D8.3	C36H8.1	C26F1.7	B0348.6b	C17G1.3	C14F5.3c
4	C30F8.2	C55C2.5b	C52A11.2	C32D5.12	F54F2.2a	C48B4.12a	C31H1.6	E03H12.2	C29F3.6	B0391.9	C26B9.6	C18A11.2
5	C34B2.11	D1007.10	C52A11.3	F29C12.1	K01G5.4	E03A3.4	C33H5.11	F02H6.4	C45H4.11	C01G10.1	C27C12.2	C23F12.3
6	C34G6.1	E01A2.6	C52E2.1	F37B1.1	K01G5.5	E03A3.5	C33H5.15	F13B12.3	C45H4.12	C04F2.1	C33D3.4	C30E1.t1
7	C36B1.12	F10D11.2	D2062.8	F39E9.5	K12H4.7b	F09F7.5c	C33H5.4	F15E6.6	C51E3.6	***C06C6.5a***	C34F6.2	C44C10.2
8	C37A2.3	F25F1.2	F13H8.8	F42A8.1	M01A8.1	F14F7.t1	E04A4.8	F20C5.5	C54D10.8	C13A2.9	C47C12.2	C52G5.2
9	C47F8.5	F35E2.7	F35C5.7	F43E2.3	M01G5.3	F26A1.10	F07C6.3	F20D12.1	C54D10.9	C13B7.2	C47C12.3	F09F9.4
10	C54C8.4	F35E2.8	F45E12.1	F44G4.5	R05D3.11	F44E2.4	F13E9.1	F20D12.6	F10A3.2	C16D9.6	F02D10.2	F39H12.3
11	D1007.3	F36H2.1	F45E12.5a	**F54C9.7**	R07E5.12	F54D8.4	**F33D4.2a**	F28F9.3	F10A3.3	C18B10.1	F15A8.4	F41E7.4
12	F02E9.7	K02F2.5	F53G2.3	F54D12.4	R10E4.7	K02F3.3	F36A4.11	F38E11.4	F11A5.10	C18G1.4a	F20B6.2	F41G4.3
13	F13G3.7	M01D7.2	F58G1.3	K02A2.1	R13A5.3	K02F3.9	F36A4.8	H04M03.2	F14F8.4	C47A10.4	F28H6.2	F47B7.7
14	F26B1.1	M04C7.2	F59E12.3	K02A2.3	R13A5.6	T04A6.3	F41A4.t1	H27C11.1	F14F8.6	F09F3.1	F28H6.3	F52D10.6
15	F30F8.3	R05D7.2	K06A1.2	K04B12.2	R151.1	T20G5.1	F52B11.4	JC8.10b	F36D3.2	F11A5.5	F41E7.8	F53A9.4
16	F35C12.3	R12E2.9	K08F8.4	K10G6.3	T04A8.15	T22F7.5	F56C4.1	K08D12.6	F45D3.1	F11A5.9	F45E6.1	F54B11.9
17	F41D3.11	T09E11.8	R05H5.2	K12H6.7	T07E3.1	Y119D3B.11	F58E2.7	K09B11.2	F45F2.1	F26F12.5	F46H5.6	F56C3.1
18	F44F1.7	T09E11.9	R06A4.6	R07C3.12	T16H12.1	Y22D7AR.12	H12I19.2	M02B7.6	F46B6.2	F37B4.5	F48D6.4b	K04G11.2
19	F47B3.1	T22A3.3	R52.1	T05H10.3	T16H12.3	Y37D8A.3	H25K10.1	T02D1.1	F56A4.4	F44E7.4b	F59F5.2	R04D3.t5
20	F47B3.7	T23G11.4	T01B7.7	T07H3.2	Y39A1A.23	Y39A1C.2	JC8.12	T12B3.4	F57B7.1b	F47B8.10	H11E01.1	R106.1
21	K10E9.1	T27F6.2	T07F8.3b	T08H4.3	Y39A1A.7	Y41C4A.13	K08B4.1a	T12E12.3	H19N07.1	F48G7.1	K03E6.1	R106.2
22	M01B12.4b	W01B11.2	T07F8.4	T15H9.4	Y48G9A.1	Y54F10BM.1	R10H10.6	T20D3.3	H19N07.2a	F52F10.2	K09A9.4	T06H11.2
23	M01B12.5	W03F11.1	T10D4.7	T15H9.5	Y50D7A.10	Y54F10BM.13	Y116A8A.7	Y38C1AA.7	K02H11.6	F59E11.10	K09A9.5	T13G4.5
24	T07D10.4	W10C8.5	T27A1.3	T27A1.2	Y56A3A.2	Y56A3A.11	Y17G9A.6	Y38C1AA.9	K12D9.4	H05B21.4	R04A9.2	T25B6.4
25	T15D6.4	Y105E8B.7	T27D12.2a	Y110A2AL.11	Y66D12A.8	Y56A3A.3	Y17G9B.5	Y39C12A.2	R09B5.4	H14N18.4a	R04B3.3	W02H3.1
26	T19B4.4	Y39G10AL.3	W02B12.12	Y17G7B.7	Y6D11A.2	Y66D12A.8	Y37E11AR.2	Y41D4B.20	R09B5.5	K07C6.4	R04D3.9	Y71H9A.1
27	T22A3.8	Y47G6A.8	W09H1.3	Y25C1A.3	Y82E9BL.1	Y6D11A.2	Y41D4A.8	Y41D4B.8	R90.3	M02H5.2	R04D3.t2	Y75D11A.t1
28	T23B3.3	Y52B11C.1	Y17G7B.12	**Y46G5A.4**	ZK328.6	Y71H2B.1	Y45F10B.1	***Y55F3AL.1***	***T01C3.10***	M162.7	R07E4.6a	

Table 7.9. **All *C. elegans* genes containing the motif TTACCGAA within their 1 kb upstream region.** Genes emboldened express in the excretory cell. Genes emboldened and italicized show expression in cell groups other than the excretory cell.

Genes containing the motif TTACCGAA/TTCGGTAA within the 1 kb upstream regions												
	Chromosome I (+)	*I (-)*	*Chromosome II (+)*	*II (-)*	*chromosome IIII (+)*	*IIII (-)*	*chromosome IV (+)*	*IV (-)*	*Chromosome V (+)*	*V (-)*	*chromosome X (+)*	*X (-)*
29	T23D8.5	ZK1053.6	Y25C1A.13	Y46G5A.31		Y75B8A.3	Y54G2A.10	Y55F3AM.14	T05B4.13	R05D8.11	R08B4.3	
30	T23D8.6		Y25C1A.2	Y48B6A.5		***Y75B8A.4***	Y55F3C.2	Y55F3AM.5	T05B4.4	R10D12.10	T06H11.2	
31	W04C9.3		Y39G8B.4	Y49F6B.8		Y82E9BR.1	Y57G11B.t1	Y62E10A.6	T05C3.4	R10D12.6	T10E10.1	
32	Y105E8A.12		Y47G7B.1	Y54G9A.5		ZK370.7	**Y62E10A.1**	Y67A10A.1	T10C6.1	R11D1.6	T13G4.1	
33	Y23H5B.9		Y49F6B.1			ZK632.12	Y73F8A.21	Y67H2A.1	T15B7.10	T08G3.8	T13G4.4	
34	Y34D9B.1a		Y51H1A.6				Y77E11A.1		T16G1.9	T10C6.4	ZC504.2	
35	Y37E3.16		Y53F4B.23				Y94H6A.9		T26H10.1	T19H12.7	**Zk455.7**	
36	Y39G10AR.17		Y53F4B.28				ZK822.2		T27C4.4b	T20D4.18		
37	Y63D3A.4		Y54G11A.7				ZK822.4		W02H5.1	T26E4.6		
38	Y63D3A.5		Y57A10B.7						W06D12.6	W02D7.10		
39	Y71F9AL.13a		Y57G7A.3						W07G4.3	Y17D7B.2		
40	Y71F9AL.9		ZC239.10						Y113G7B.8	Y38C9B.1		
41	Y71G12B.27		ZC239.19						Y113G7B.9	Y43F8C.5		
42	Y71G12B.5		ZK1320.9						Y17D7A.4	Y45G12C.10		
43	Y76G2A.2		ZK666.6						Y19D10B.4	Y45G12C.6		
44	ZK770.1		ZK84.4						Y20C6A.1	Y50E8A.14		
45			ZK930.2						Y20C6A.2	Y51A2D.5		
46									Y39B6A.12	ZK1005.1		
47									Y39B6A.15			
48									Y39D8B.1			
49									Y45G12C.1			
50									Y45G12C.3			
51									Y46H3D.7			
52									Y49G5B.1			
53									Y60C6A.1			
54									Y60C6A.t1			
55									ZC196.1			
56									ZC513.2			
57									ZK262.10			
58									ZK856.6			

Table 7.9 continued. All *C. elegans* genes containing the motif TTACCGAA within their 1 kb upstream region. Genes emboldened express in the excretory cell. Genes emboldened and italicized show expression in cell groups other than the excretory cell.

7.10: Discussion.

In Chapter 6 a candidate *cis*-acting element, which potentially directs gene expression in the excretory cell, was identified from the 2 kb excretory set of *C. elegans*. Before this candidate element could be verified by *in vivo* experimentation it was considered important to obtain additional *in silico* evidence of it having a role as a regulatory element. This was achieved by looking for the presence of the element in two different data sets, one data set consisting of genes expressing in the excretory cell (but not used in any previous analyses) and the second data set consisting of *C. briggsae* orthologues of *C. elegans* excretory cell-expressing genes. Analyses of these two test sets revealed that the highest scoring motif (TTACCGAA) from the 2 kb excretory set was also present in both the excretory and *C. briggsae* test sets. However, the element occurred only once (out of 26 genes) in the *C. briggsae* set. The apparent low occurrence of the motif in the *C. briggsae* test set may be attributed to the fact that the *C. briggsae* and the *C. elegans* genome diverged approximately 25-50 million years and it is therefore possible that the candidate *cis*-acting element may have subtly diverged (while remaining functional) between the two genomes over this time period. In this analyses only exact matches were being detected therefore a single base change will result in a failure to detect the motif in the *C. briggsae* genome. In addition, the position of the motif between the two genomes may not be strictly conserved and the element may be present in the *C. briggsae* orthologues but beyond the region being analysed.

Exact matches of the candidate element was detected in four of the 23 genes in the excretory test set suggesting that the element was not an absolute requirement for expression in the excretory cell. However, the motif was located within the 1 kb upstream region of two genes from the excretory test set and as well as in the gene from the *C. briggsae* test set suggesting that the motif may be a *cis*-acting element. In addition, other variations of this motif may have been detected if degeneracy was allowed.

Searching the *C. elegans* genome with the TTACCGAA element revealed that it occurred on average once every 27 kb of DNA sequence. In the positive sets of the 2 kb excretory (Chapter 6) and excretory test sets it occurs 5 in 40 kb and 4 in 46 kb, respectively,

indicating a degree of enrichment in the data sets. Predictably the motif was underrepresented in the negative set where it occurred once per 268 kb of DNA.

To conclude motif TTACCGAA is considered a good candidate *cis*-acting element. However, experimental validation is required to definitively establish that the motif is a true *cis*-acting element that is capable of directing expression in the excretory cell. The final chapter of this thesis will discuss a number of different approaches that could be used to examine the regulatory element-status of this motif in the *in vivo* setting (Section 8.2.1).

Chapter 8

8: Final discussion and future work.

8.1: Summary of results.

8.1.1: Evaluation of gene expression pattern data.

Gene expression patterns provide important information regarding the temporal and spatial control of genetic information, thereby providing a four-dimensional model of organism development (Hope *et al.*, 1996). For the nematode *C. elegans* three main experimental approaches have been used to generate gene expression pattern data. These approaches are: reporter gene fusion (with GFP or lacZ) providing information of promoter activity, mRNA *in situ* hybridization identifying the distribution of gene-specific transcripts, and immunostaining identifying the specific distribution of protein.

Although there are advantages and disadvantages to all three approaches, particular criticism has been directed at the reporter gene fusion approach, suggesting that activity of a promoter in directing expression of a marker protein (e.g. GFP or β-galactosidase) may not truly reflect the expression of an endogenous gene (Fire, 1995). The integrity and reliability of gene expression patterns produced by the reporter gene fusion approach is of particular importance to this project. Therefore at the commencement of this research it was essential to evaluate the consistency of data by performing a comparison of text descriptions generated by the three different approaches of producing gene expression patterns. At the start of the project there were approximately 240 *C. elegans* gene expression patterns in the ACeDB database (version WS8, July 1999). However, it was known that this database did not contain a comprehensive catalogue of all the determined expression patterns, and there were many additional expression patterns published in scientific journals. Therefore a systematic survey of the literature was performed which resulted in the identification of a further 202 gene expression patterns. This information was then submitted for inclusion into WormBase/ACeDB and a comprehensive analysis of all gene expression patterns in the updated WormBase database was performed. The salient findings from these studies were as follows:

- Approximately 30% of the gene expression patterns generated using the three main approaches were identical.

- Greater than a third of reporter gene fusion patterns are identical to those produced by immunostaining with the remainder having a high degree of similarity and only minor discrepancies restricted to certain cell groupings.
- Some discrepancies were believed to reflect the different points at which gene expression can be controlled e.g. the post transcriptional level.
- Many of the differences were due to a lack of detection of germline expression, a phenomenon that is known to occur with reporter genes (Kelly *et al.*, 1997; Kelly *et al.*, 1998).
- The resolution obtained from the reporter gene fusion analyses was of a superior quality to that obtained from the mRNA *in situ* hybridization approach.
- When compared to mRNA *in situ* hybridization, reporter gene fusion and immunostaining approaches are more sensitive for detecting low levels of expression.

Although there were no major discrepancies between the gene expression patterns produced by the different experimental approaches, immunostaining and reporter gene fusion techniques provided the most detailed description of *C. elegans* gene expression. Importantly, data generated by the reporter gene fusion techniques was reliable and consistent with that produced by other experimental approaches.

8.1.2: Computational detection of *cis*-acting regulatory elements.

The recent availability of the complete nucleic acid sequence of genomes of a number of different organisms, including *C. elegans* has provided new opportunities to study gene regulation. One such opportunity, and that has been pursued here, is the identification of DNA motifs that are common to the promoter sequences of genes which share similar gene expression patterns. If a motif is specific to the promoters of putatively co-regulated genes (i.e. genes having similar expression patterns) it is possible that the motif is a *cis*-acting regulatory element required for appropriate gene expression (i.e. spatial, temporal and level of expression). The analyses of expression patterns (described above) resulted in the identification of 20 genes that are expressed in one particular *C. elegans* cell, the excretory cell. It is considered probable that many of the 20 genes expressed in the excretory cell are potentially co-regulated and therefore contain the same *cis*-acting regulatory elements. The

excretory cell of *C. elegans* was considered to be particularly suitable for the study of gene expression and the identification of *cis*-acting elements because it is the simplest tissue, differentiated as a single cell. Therefore the mechanisms controlling gene expression may be less complex than in other cell types.

In order to identify *cis*-acting regulatory elements, the region of these genes able to drive this expression pattern as demonstrated by the reporter gene fusion approach was analyzed using two software packages, MEME [http://meme.sdsc.edu/meme/website/] and SPEXS [http://ep.ebi.ac.uk/EP/SPEXS/].

8.1.2.1: Analysis of the promoter region using the MEME software.

The MEME output contained many motifs that were common to the promoter regions of a number of genes that express in the excretory cell. However many of the motifs were of low sequence complexity and highly repetitive (e.g. TTTTTTTTTTTT, AAAAAAAAAAAA and AAAAATAAAAAA), and therefore unlikely to be *cis*-acting elements thus raising questions about the value of the approach for motif detection in *C. elegans*. Although MEME has previously been reported to successfully identify *cis*-acting elements from the promoter regions of co-regulated genes in unicellular organisms such as yeast, there appeared to be limitations in its use for the analysis of genes from more complex, multicellular organisms. This may be because the genome of multicellular organisms is less densely organized (Mewes *et al.*, 1997, Comeron, 2001; Poole *et al.*, 2003) and often contains many repetitive elements that hinder the detection of candidate *cis*-acting elements by MEME, particularly with the TCM model (which allows for multiple occurrences of a motif in a DNA sequence). In addition, a further limitation to the usefulness of MEME is that the programming of the algorithm is complex and convoluted and as a consequence it is difficult to modify such that a more meaningful output can be generated.

8.1.2.2: Analysis of the promoter region using the SPEXS software.

In contrast to MEME, the output from SPEXS was much simpler (only two pieces of information provided, motif and frequency of occurrence) although a vast number of motifs (several thousands) were detected. Based solely on the SPEXS output it was difficult to

assign a priority to each motif in terms of the likelihood of each being a *cis*-acting element. This was because SPEXS detects DNA motifs simply based on sequence without considering biological factors that may make a motif more, or less likely, to be a regulatory element. Therefore a scoring strategy was devised that incorporated different weightings for various biological factors (frequency, position and DNA sequence complexity) that are likely to influence whether a motif is a *cis*-acting element. As a consequence motifs that occurred with high frequency, within the 1 kb region upstream from the translational start, and with high DNA sequence complexity, were assigned a greater score as these motifs are the more likely candidates for *cis*-acting regulatory elements.

For motifs that were identical in sequence but varied in length, a proportion of the Total Score for the shorter motif was added to the score of the larger motif, and vice versa. This allowed for degeneracy at the end position of the motif and provided an additional weighting for motifs that were very similar in sequence. Furthermore, scores for motifs that differed by a single nucleotide at an internal position but were otherwise identical were mutually increased by an appropriate factor. This strategy relaxed the rigidity of SPEXS, which identifies motifs only of exact matches, and allowed a degree of degeneracy which is an important characteristic of *cis*-acting elements.

Initially the SPEXS software was executed with the complete genomic DNA sequences (of the fragments used in reporter gene fusion experiments) for genes able to drive expression in the excretory cell. The output from this analysis was used to develop the scoring strategy. In order to assess the effectiveness of the strategy it was evaluated with the 2 kb upstream regions of genes from a *C. elegans* muscle gene set for which a *cis*-acting element had previously been characterized by Guhatharkurta *et al.* (2002b). The use of the scoring strategy resulted in the successful identification of the previously identified muscle element.

Following the testing of the scoring strategy on the *C. elegans* muscle data set, it was noticeable that the data set consisting of the entire genomic insert (from excretory cell expressing genes) yielded a greater number of motifs than the muscle set. This was probably due to a greater level of noise (non-specific motifs) because of the larger size of

the fragment of DNA being evaluated. In addition, the detection of the muscle element within the 2 kb upstream region suggested that a *cis*-acting element mediating excretory cell expression may also be present within this region of the promoters of the respective *C. elegans* gene. Therefore the analysis was repeated with the excretory set using 2 kb of DNA sequence from the upstream region of each gene. The highest scoring motif and most likely candidate for a *cis*-acting element was TTACCGAA. However before this motif could be verified experimentally (*in vivo* laboratory studies) further evidence was sought by examining the frequency of this motif on two additional test data sets.

The first test set consisted of 23 *C. elegans* genes that expressed in the excretory cell but had not been used in the previous analyses. The second test set consisted of 26 *C. briggsae* orthologues of the *C. elegans* genes which express in the excretory cell. The highest scoring motif TTACCGAA from the 2 kb excretory set was also detected, although in fewer genes, in the excretory test and the *C. briggsae* test set. This suggested that the motif TTACCGAA was a good candidate for a *cis*-acting element mediating expression in the excretory cell.

The *C. elegans* genome was then searched with the motif TTACCGAA. The motif was found to occur on 3771 occasions with 447 containing the motif within their 1 kb upstream regions. Of these 447 genes that contained the motif, five were known to express in the excretory cell and had been used in the earlier SPEXS analyses, four genes did not express in the excretory cell and the remaining 438 genes were novel and expression patterns were not available. After examining WormBase gene ontology data available for 150 of the 447 genes, it was apparent that the majority of these genes played a role in transport (cation, chloride, hydrogen and sodium transport). This finding is consistent with the known functions of the excretory cell, i.e. osmoregulation, secretion and export of hormones to target tissues and excretion of metabolic waste (Nelson *et al.*, 1983; Golden and Riddle, 1982; Bürglin and Ruvkun, 2001; Strange, 2003). It is therefore entirely possible, that many if not the majority, of the 438 novel genes which contain the element in the 1 kb upstream region show expression in the excretory cell. The observation that the genes containing the motif in their respective promoter regions have functions that are compatible

with the role of the excretory cell further suggests that the motif is a *cis*-acting element which directs expression in the excretory cell.

8.2: Future work and implications.

8.2.1: *In vivo* testing of candidate motifs.

Application of the scoring strategy based on biological variables to the SPEXS output resulted in the identification of a number of potential *cis*-acting elements, with the motif, TTACGGAA considered the most likely element. To provide additional evidence that a candidate element such as TTACGGAA is a true regulatory element it is necessary to further evaluate the motif in question. One approach, although not definitive, of examining whether a motif is a true *cis*-acting element is to select a number of genes at random from the list of predicted genes which contain the candidate element and characterize their respective expression patterns. If a sufficiently high number of genes express in the excretory cell (when compared to genes which do not contain the element) it is likely that the element in question is required for excretory cell expression.

To conclusively demonstrate that this motif is indeed a regulatory element it must be formally tested in *in vivo* assays. There are several different approaches that can be employed to verify that the element identified is required for specific expression. One such approach involves site directed mutagenesis and thereby elimination of the element from the promoter region of excretory cell-expressing genes. If the motif is an element essential to expression of the gene in the excretory cell, clearly removal of the element should result in the absence of expression when analysed by reporter gene fusion approaches.

Alternatively, the candidate element can be cloned into the promoter region of a gene fusion construct of a promoter previously characterized, and known to express in *C. elegans*, but in cells other than the excretory cell. For example the candidate element can be cloned into a basal promoter such as the *pes-10* element (Kirouac and Sternberg, 2003). The expression pattern of promoter and newly inserted candidate element can then be characterized by the reporter gene fusion approach. If the candidate element is a true *cis*-acting regulatory element then expression should be evident in the excretory cell.

If a candidate *cis*-acting element can be validated *in vivo* the strategy adopted in this thesis will be vindicated. Following such confirmatory experiments a next step would be to make the software fully-automated. This would require executing SPEXS and performing all subsequent analyses necessary for the scoring strategy as a single step such that the user will only be obligated to provide the input sequences. The final output would then consist of motifs ranked based on the scores calculated by the algorithm.

8.2.2: Incorporating biological characteristics into an algorithm.

When used alone, neither the enumerative (SPEXS) nor the alignment (MEME) approaches were successful in the identification of *cis*-acting elements. The usefulness of these algorithms is believed to be compromised by random stretches of simple subsequences or the presence of *cis*-acting elements that regulate expression in cell types other than the excretory cell. However, the application of a scoring strategy based on probable biological features of elements to the output generated by SPEXS resulted in the identification of a number of potential candidate *cis*-acting elements, in particular the motif, TTACCGAA. Therefore it would appear essential that the predictive capability of any fully-automated approach designed to identify elements will be poor unless biological knowledge of regulatory elements is incorporated into the algorithm.

The incorporation of any biological features into an algorithm or scoring strategy clearly requires a substantial understanding of the characteristics of *cis*-acting elements. Unfortunately, as yet, few elements have been fully characterized and it is therefore unclear as to what biological factors are of key importance to the functioning of an element. In this study a "weighting" was assigned to biological factors (frequency, position and sequence complexity) that are thought to be associated with elements. In an attempt to improve specificity of detection of elements, the relative emphasis on each of these weightings could be refined to find an optimal weighting for each factor. A machine learning algorithm such as decision trees may be used for such a purpose by analyzing a data set that contains a previously characterized *cis*-acting element (Quinlan *et al.*, 1993). The machine learning algorithm would then identify a set of rules which could then be applied to search for unknown *cis*-acting elements from co-regulated genes. However this is not a trivial

undertaking because a large number of previously characterized *cis*-acting elements are required, are not yet available.

The level of expression of a particular gene may be one additional biological variable that can be incorporated into any method for identifying elements. This is because it has been shown that the higher the level of expression the greater the number of a particular *cis*-acting element in the promoter of a gene (Gaudet and Mango, 2002). However the level of expression of a particular gene may be difficult to quantitate and lead to inaccuracies.

In addition to refining the weighting applied to each biological variable it may also be beneficial to modify how each biological factor is measured. For example as there is some evidence (Kirouac and Sternberg, 2003) that elements are typically found in close proximity to the translational start site, the scoring strategy placed an emphasis on the position of the motif within the promoter region and motifs were characterized as either greater than, or less than 1 kb from the translational start site. In any future scoring strategy (or incorporated into a fully-automated algorithm) a possible modification to the variable of motif position may involve a more precise description (i.e. the coordinates of the motif in relation to the translational start site) rather than simply being placed in either of the two categories described above. This modification may result in a greater specificity of element identification.

A further point that must be considered when refining a strategy to identify regulatory elements concerns the fact that a number of transcription factors function as multimers and bind to more than one element (Drees *et al.*, 1997). Detection of motifs such as these will also require any algorithm to allow for distinct gaps between the motifs. The detection of elements can be even further complicated if the transcription factor binds in the manner of the zinc finger transcription factor which bind to seven different elements (Clark and Berg, 1998). Each of these elements would have to be identified and tested experimentally as the absence of a single element may lead to weak or no expression.

8.2.3: Future applications.

It is anticipated that ultimately a fully-automated strategy for detecting *cis*-acting elements will be developed. Such an approach could then be applied to gene expression data sets from more complex genomes including, of course, the human. However, due to the large amount of junk DNA, interpretation of the human genome sequence with the aim of identifying regulatory elements may prove to be far more problematic than for *C. elegans*. Therefore to increase the specificity of detection of motifs amongst the vastness of the junk DNA there is potentially an even greater requirement for a better appreciation of the biological characteristics of *cis*-acting elements.

The ability to understand the regulatory mechanisms of human gene expression can have numerous long-term applications. This information will provide a greater understanding of cellular and genetic processes, and may ultimately be used to gain a better appreciation of pathological mechanisms. Such pathologies may include cancer, a disease invariably associated with aberrant gene expression (Jubb *et al.*, 2003). In the longer-term understanding how the cancer-controlling, oncogenes or tumour suppressor genes are expressed may lead to the identification of novel sites of therapeutic intervention which may be exploited to improve the treatment of this life-threatening disease.

References

References.

Adams M. D., Celniker S. E., Holt R. A., Evans C. A., Gocayne J. D., Amanatides P. G., Scherer S. E., Li P. W., Hoskins R. A., Galle R. F., George R. A., Lewis S. E., Richards S., Ashburner M., Henderson S. N., Sutton G. G., Wortman J. R., Yandell M. D., Zhang Q., Chen L. X., Brandon R. C., Rogers Y. H., Blazej R. G., Champe M., Pfeiffer B. D., Wan K. H., Doyle C., Baxter E. G., Helt G., Nelson C. R., Gabor G. L., Abril J. F., Agbayani A., An H. J., Andrews-Pfannkoch C., Baldwin D., Ballew R. M., Basu A., Baxendale J., Bayraktaroglu L., Beasley E. M., Beeson K. Y., Benos P. V., Berman B. P., Bhandari D., Bolshakov S., Borkova D., Botchan M. R., Bouck J., Brokstein P., Brottier P., Burtis K. C., Busam D. A., Butler H., Cadieu E., Center A., Chandra I., Cherry J. M., Cawley S., Dahlke C., Davenport L. B., Davies P., de Pablos B, Delcher A., Deng Z, Mays A. D., Dew I, Dietz S. M., Dodson K., Doup L. E., Downes M., Dugan-Rocha S., Dunkov B. C., Dunn P., Durbin K. J., Evangelista C. C., Ferraz C., Ferriera S., Fleischmann W., Fosler C., Gabrielian A. E., Garg N. S., Gelbart W. M., Glasser K., Glodek A., Gong F., Gorrell J. H., Gu Z., Guan P., Harris M., Harris N. L., Harvey D., Heiman T. J., Hernandez J. R., Houck J., Hostin D., Houston K. A., Howland T. J., Wei M. H., Ibegwam C., Jalali M., Kalush F., Karpen G. H., Ke Z., Kennison J. A., Ketchum K. A., Kimmel B. E., Kodira C. D., Kraft C., Kravitz S., Kulp D., Lai Z., Lasko P., Lei Y., Levitsky A. A., Li J., Li Z., Liang Y., Lin X., Liu X., Mattei B., McIntosh T. C., McLeod M. P., McPherson D., Merkulov G., Milshina N. V., Mobarry C., Morris J., Moshrefi A., Mount S. M., Moy M., Murphy B., Murphy L., Muzny D. M., Nelson D. L., Nelson D. R., Nelson K. A., Nixon K., Nusskern D. R., Pacleb J. M., Palazzolo M., Pittman G. S., Pan S., Pollard J., Puri V., Reese M. G., Reinert K., Remington K., Saunders R. D., Scheeler F., Shen H., Shue B. C., Siden-Kiamos I., Simpson M., Skupski M. P., Smith T., Spier E., Spradling A. C., Stapleton M., Strong R., Sun E., Svirskas R., Tector C., Turner R., Venter E., Wang A. H., Wang X., Wang Z. Y., Wassarman D. A., Weinstock G. M., Weissenbach J., Williams S. M., WoodageT., Worley K. C., Wu D., Yang S., Yao Q. A., Ye J., Yeh R. F., Zaveri J. S., Zhan M., Zhang G., Zhao Q., Zheng L., Zheng X. H., Zhong F. N., Zhong W., Zhou X., Zhu S., Zhu X., Smith H. O., Gibbs R. A., Myers E. W., Rubin G. M. and Venter J. C. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287:** 2185-95.

Ahringer J. (1996) Posterior patterning by the *Caenorhabditis elegans* even-skipped homolog *vab-7*. *Gene Dev.*, **10:** 1120-1130.

Albertson D. G. and Thomson J. N. (1976) The pharynx of *Caenorhabditis elegans*. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **275:** 299-325.

Altschul S. F., Gish W., Myers E. W. and Lipman D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215:** 403-410.

Alwine J. C., Kemp D. J. and Stark G. R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U S A*, **74:** 5350-5354.

Ashrafi K., Chang F. Y., Watts J. L., Fraser A. G., Kamath R. S., Ahringer J. and Ruvkun G. (2003) Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes. *Nature*, **421:** 268-272.

Ayala F. J., Barrio E. and Kwiatowski J. (1996) Molecular clock or erratic evolution? A tale of two genes. *Proc. Natl. Acad. Sci. USA*, **93:** 11729-11734.

Bailey T. L. And Elkan C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Intell. Sys. Mol. Biol.*, **2:** 28-36.

Bailey T. L. and Elkan C. (1995a) Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Machine Learning Journal*, **21:** 51-83.

Bailey T. L. and Elkan C. (1995b) The value of prior knowledge in discovering motifs with MEME. *Intell. Sys. Mol. Biol.*, **3:** 21-29.

Bailey T. L., Baker M. E. and Elkan C. P. An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. *J. Steroid Biochem. Mol. Biol.*, **62:** 29-44.

Baugh L. R., Hill A. A., Brown E. L. and Hunter C. P. (2001) Quantitative analysis of mRNA amplification by *in vitro* transcription. *Nucleic Acids Res.*, **29:** E29.

Berkowitz L.A. and Strome S. (2000) MES-1, a protein required for unequal divisions of the germline in early *C. elegans* embryos, resembles receptor tyrosine kinases and is localized to the boundary between the germline and gut cells. *Development,* **127:** 4419-31.

Berks M. and the *C. elegans* genome mapping and sequencing consortium. (1995) The *C. elegans* genome sequencing project. *Genome Res.*, **5:** 99-104.

Birchall P. S., Fishpool R. M. and Albertson D. G. (1995) Expression patterns of predicted genes from the *C. elegans* genome sequence visualized by FISH in whole organisms. *Nat. Genet.*, **11**: 314-320.

Blanchette M. and Tompa M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**: 739-748.

Blaxter M. (1998) *Caenorhabditis elegans* is a nematode. *Science*, **282:** 2041-2046.

Brazma A., Jonassen I., Vilo J. and Ukkonen E. (1998a) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, **8:** 1202-1215.

Brazma A., Jonassen I., Vilo J. and Ukkonen E. (1998b) Pattern discovery in biosequences. *In Proceedings of the Fourth International Colloquitum on Grammar Inference, Lecture Notes in Artificial Intelligence, Springer, New York, NY*, **1433:** 257-270.

Brazma A., Jonassen I., Eidhammer I. and Gilbert D. (1998c) Approaches to the automatic discovery of patterns in biosequences. *J. Comput. Biol.*, **5:** 279-305.

Brazma A. and Vilo J. (2000) Gene expression data analysis. *FEBS Lett.*, **480:** 17-24.

Bucher P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, **9:** 400-407.

Buechner M. (2002) Tubes and the single *C. elegans* excretory cell. *Trends Cell Biol.*, **12**: 479-484.

Bürglin T. R. and Ruvkun G. (2001) Regulation of ectodermal and excretory function by the C. elegans POU homeobox gene *ceh-6*. *Development*, **128:** 779-790.

Bussemaker H. J., Li H. and Siggia E. D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27:** 167-71.

Chalfie M. (1998) The worm revealed. *Nature*, **396**: 620-621.

Chen S., Zhou S., Sarkar M., Spence A. M. and Schachter H. (1999) Expression of three *Caenorhabditis elegans* N-acetylglucosaminyltransferase I genes during development. *J. Biol. Chem.*, **274:** 288-297.

Cherry J. M., Adler C., Ball C., Chervitz S. A., Dwight S. S., Hester E. T., Jia Y., Juvik G., Roe T., Schroeder M., Weng S. and Botstein D. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26:** 73-9.

Chisholm A. D. and Jin Y. (2001) *Caenorhabditis elegans* as an experimental organism. In Encyclopaedia of life sciences. *Nature Publishing group*, 1-7.

Cho J. H., Oh Y. S., Park K. W., Yu J., Choi K. Y., Shin J. Y., Kim D. H., Park W. J., Hamada T., Kagawa H., Maryon E. B., Bandyopadhyay J. and Ahnn J. (2000) Calsequestrin, a calcium sequestering protein localized at the sarcoplasmic reticulum, is not essential for body-wall muscle function in *Caenorhabditis elegans*. *J. Cell Sci.*, **113:** 3947-58.

Clarke N. D and Berg J. M. (1998) Zinc Fingers in *Caenorhabditis elegans*: Finding Families and Probing Pathways. *Science*, **282:** 2018-22.

Claverie J. M. and audic S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12:** 431-439.

Cliften P., Hillier L., Fulton L., Graves T., Miner T., Gish W., Waterson R. and Johnstone M. (2001) Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.*, **11:** 1175-1186.

Comeron J. M. (2001) What controls the length of noncoding DNA? *Curr. Opin. Genet. Dev.*, **11:** 652-9.

Coulson A., Sulston J., Brenner S. and Karn J. (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U S A*, **83:** 7821-7825.

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

Coulson A., Kozono Y., Lutterbach B., Shownkeen R., Sulston J., Waterston R. (1991) YACs and the *C. elegans* genome. *Bioessays*, **13:** 413-417.

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.

Culetto E., Combes D., Fedon Y., Roig A., Toutant J. P. and Arpagaus M. (1999) Structure and promoter activity of the 5' flanking region of ace-1, the gene encoding acetylcholinesterase of class A in *Caenorhabditis elegans*. *J. Mol. Biol.*, **290:** 951-66.

Dalley B. K., Rogalski T. M., Tullis G. E., Riddle D. L. and Golomb M. (1993) Posttranscriptional regulation of RNA polymerase-II in *Caenorhabditis elegans*. *Genetics*, **133:** 237-245.

Davidson D., Bard J., Kaufman M. and Baldock R. (2001) The MouseAtlas Database: a community resource for mouse development. *Trends Genet.*, **17:** 49-51.

Deppe U., Schierenberg E., Cole T., Krieg C., Schmitt D., Yoder B. and von Ehrenstein G. (1978) Cell lineages of the embryo of the nematode *C. elegans*. *Proc. Natl. Acad. Sci. USA*, **75:** 376-380.

DeRisi J. L., Vishwanath R., Iyer R. and Brown P. O. (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, **278**: 680-686.

Drees B. L., Grotkopp E. K. and Nelson H. C. M. (1997) The GCN4 leucine zipper can functionally substitute for the heat shock transcription factor's trimerization domain. *J. Mol. Biol.*, **273**: 61-74.

Dufourcq P., Chanal P., Vicaire S., Camut E., Quintin S., den Boer B. G., Bosher J. M. and Labouesse M. (1999) lir-2, lir-1 and lin-26 encode a new class of zinc-finger proteins and are organized in two overlapping operons both in *Caenorhabditis elegans* and in *Caenorhabditis briggsae*. *Genetics*, **152:** 221-35.

Duret L. and Bucher P. (1997) Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.*, **7:** 399-405.

Edgar L.G., Carr S., Wang H. and Wood W. B. (2001) Zygotic expression of the caudal homolog pal-1 is required for posterior patterning in *Caenorhabditis elegans* embryogenesis. *Dev. Biol.*, **229:** 71-88.

Egan C. R., Chung M. A., Allen F. L., Heschl M. F., Van Buskirk C. L. and McGhee J. D. (1995) A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans* ges-1 gene centers on two GATA sequences. *Dev. Biol.*, **170:** 397-419.

Eisen M.B., Spellman P.T., Brown P.O., Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A*, **95**: 14863-14868.

Evans T. C., Crittenden S. L., Kodoyianni V. and Kimble J. (1994) Translational control of maternal glp-1 mRNA establishes an asymmetry in the *C. elegans* embryo. *Cell*, **77:** 183-94.

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

Fire A. (1995) Fire Lab. Vector kit.

Flybase Consortium (1998) FlyBase: a Drosophila database. *Nucleic Acids Res.*, **26:** 85-88.

Fraser A. G., Kamath R. S., Zipperlen P., Martinez-Campos M., Sohrmann M. and Ahringer J. (2000) Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature*, **408:** 325-30.

Frech, K., Herrmann G. and Werner, T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.*, **21:** 1655-1664.

Frith M. C., Hansen U. and Weng Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17:** 878-889.

Fukushige T. and Siddiqui S. S. (1995) Effect of the dpy-20 and rol-6 cotransformation markers on alpha-tubulin gene expression in *C. elegans* transformants. *Transgenic Res.*, **4:** 332-40.

Fukushige T., Hawkins M. G. and McGhee J. D. (1998) The GATA-Factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. *Dev. Biology,* **198**: 286-302.

Galas D. J. Eggert M. and Waterman M. S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli. J. Mol. Biol.*, **186:** 117-128.

Gaudet J. P., Van der Elst I. and Spence A. M. (1996) Post-transcriptional regulation of sex determination in *Caenorhabditis elegans* – Widespread expression of the sex-determining gene fem-1 in both sexes. *Mol. Biol. Cell*, **7:** 1107-1121.

Gaudet J. and Mango S. E. (2002) Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science*, **295:** 821-5.

Giegerich R. and Kurtz S. (1995) A comparison of imperative and purely functional suffix tree constructions. *Sci. Comput. Programming*, **25:** 187-218.

Giegerich R. and Kurtz S. (1997) From Ukkonen to McCreight and Weiner: A unifying view of linear-time suffix tree construction. *Algorithmica*, **19:** 331-353.

Gilleard J. S., Barry D. and Johnstone I. L. (1997) *cis* regulatory requirements for hypodermal cell-specific expression of the *Caenorhabditis elegans* cuticle collagen gene *dpy-7*. *Mol. Cell. Biol.*, **17:** 2301-2311.

Golden J W. and Riddle D. L. (1982) A pheromone influences larval development in the nematode *Caenorhabditis elegans*. *Science*, **218:** 578-580.

Gomez-Escobar N., Gregory W. F., Britton C., Murray L., Corton C., Hall N., Daub J., Blaxter M. L. and Maizels R. M. (2002) Abundant larval transcript-1 and -2 genes from Brugia malayi: diversity of genomic environments but conservation of 5' promoter sequences functional in *Caenorhabditis elegans*. *Mol. Biochem. Parasitol.*, **125:** 59-71.

Gower N. J., Temple G. R., Schein J. E., Marra M., Walker D. S. and Baylis H. A. (2001) Dissection of the promoter region of the inositol 1,4,5-trisphosphate receptor gene, itr-1, in *C. elegans*: a molecular basis for cell-specific expression of IP3R isoforms. *J. Mol. Biol.*, **306:** 145-57.

Grabe N. (2000) AliBaba2: Context Specific Identification of Transcription factor Binding Sites. *In Silico Biol.*, **1:** 0005 [http://www.bioinfo.de/isb/2000/01/0019/].

Graham P. L., Johnson J. J., Wang S., Sibley M. H., Gupta M. C. and Kramer J. M. (1997) Type IV collagen is detectable in most, but not all, basement membranes of *Caenorhabditis elegans* and assembles on tissues that do not suppress it. *J. Cell Biol.*, **137:** 1171-1183.

Grant R. A., Rould M. A., Klemm J. D. and Pabo C. O. (2000) Exploring the role of glutamine 50 in the homeodomain-DNA interface: crystal structure of engrailed (Gln50 --> ala) complex at 2.0 A. *Biochemistry*, **39:** 8187-92.

Gregoire F. M., Chomiki N., Kachinskas D. and Warden C. H. (1998) Cloning and developmental regulation of a novel member of the insulin-like gene family in *Caenorhabditis elegans*. *Biochem. Biophys. Res. Commun.*, **249:** 385-90.

Grillo M. and Margolis F.L. (1990) Use of reverse transcriptase polymerase chain reaction to monitor expression of intronless genes. *Biotechniques*, **9:** 262, 264, 266-8.

Grundy W. N., Bailey T. L. and Elkan C. P. (1996) ParaMEME: a parallel implementation and a web interface for DNA and protein motif discovery tool. *Comput. Appl. Biosci.*, **12:** 303-310.

Guhathakurta D., Palomar L., Stormo G. D., Tedesco P., Johnson T. E., Walker D. W., Lithgow G., Kim S. and Link C. D. (2002a) Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome res.*, **12:** 701-712.

Guhathakurta D., Palomar L., Hresko M. C., Waterston R. H. and Stormo G. D. (2002b) Identifying muscle regulatory elements and genes in the nematode *Caenorhabditis elegans*. *Pac. Symp. Biocomput.*, 425-436.

Haerry T. E. and Gehring W. J. (1996) Introns of the mouse Hoxa-7 gene contains conserved homeodomain binding sites that can function as an enhancer element in *Drosophila*. *Proc. Natl. Acad. Sci. USA*, **93:** 13884-13889.

Halfon M. S., Grad Y., Church G. M. and Michelson A. M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12:** 1019-1028.

Hall D. H. and Russell R. L. (1991) The posterior nervous system of the nematode *Caenorhabditis elegans*: Serial reconstruction of identified neurons and complete pattern of synaptic interactions. *J. Neurosci.*, **11:** 1-22.

Hardison R. C., Oeltjen J. and Miller W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.*, **7:** 959-966.

Harfe B. D. and Fire A. (1998) Muscle and nerve-specific regulation of a novel NK-2 class homeodomain factor in *Caenorhabditis elegans*. *Development*, **125:** 421-429.

Harris T. W., Lee R., Schwarz E., Bradnam K., Lawson D., Chen W., Blasier D., Kenny E., Cunningham F., Kishore R., Chan J., Muller H-M., Petcherski A., Thorisson G., Day A., Bieri T., Rogers A., Chen C-K., Spieth J., Sternberg P., Durbin R. and Stein L. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31:** 133-137.

Hayashizaki S., Iino Y. and Yamamoto M. (1998) Characterization of the C. elegans gap-2 gene encoding a novel Ras-GTPase activating protein and its possible role in larval development. *Genes Cells*, **3:** 189-202.

Hermann G. J., Leung B. and Priess J. R. (2000) Left-right asymmetry in *C. elegans* intestine organogenesis involves a LIN-12/Notch signaling pathway. *Development*, **127:** 3429-40.

Hertz G. Z., Hartzell, III, G. W., and Stormo G. D. (1990) Identification of consensus patterns in inaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6:** 81-92.

Hertz G. Z., and Stormo G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15:** 563-577.

Heyer L. J., Kruglyak S. and Yooseph S. (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes. *Genome Res.*, **9:** 1106-1119.

Hill A. A., Hunter C. P., Tsung B. T., Tucker-Kellogg G. and Brown E. L. (2000) Genomic analysis of gene expression in *C. elegans*. *Science*, **290:** 809-812.

Hishiki T., Kawamoto S., Morishita S. and Okubo K. (2000) BodyMap: a human and mouse gene expression database. *Nucleic Acids Res.*, **28:** 136-138.

Hodgkin J., Plasterk R. H. A. and Waterston R. H. (1995) The nematode *Caenorhabditis elegans* and its genome. Science, **270:** 410-414.

Hodgkin J. (2001) What does a worm want with 20,000 genes? *Genome Biol.*, **2:** 2008.1-2008.4.

Hoier E. F., Mohler W. A., Kim S. K. and Hajnal A. (2000) The *Caenorhabditis elegans* APC-related gene apr-1 is required for epithelial cell migration and Hox gene expression. *Genes Dev.*, **14:** 874-86.

Holter N. S., Mitra M., Martin A., Cieplak M., Banavar J. R. and Fedoroff N.V. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Acad. Natl. Sci. USA*, **97:** 8409-8414.

Hope I.A., Albertson D.G., Martinelli S.D., Lynch A.S., Sonnhammer E., Durbin R. (1996) The *C. elegans* expression pattern database: a beginning. *Trends Genet.*, **12:** 370-371.

Horten P. B. and Kanehisa M. (1992) An assessment of neural network and statistical approaches for prediction of *E.coli* promoter sites. *Nucleic Acids Res.*, **20:** 4331-4338.

Hughes J. D., Estep P. W. Tavazoie S. and Church G. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces Cerevisiae. J. Mol. Biol.*, **296:** 1205-1214.

Hunter C. P. and Kenyon C. (1996) Spatial and temporal controls target pal-1 blastomere-specification activity to a single blastomere lineage in *C. elegans* embryos. *Cell*, **87:** 217-26.

Jareborg N., Birney E. and Durbin R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9:** 815-824.

Janning W. (1997) FlyView, a Drosophila image database, and other Drosophila databases. *Semin. Cell Dev. Biol.*, **8:** 469-475.

Jiang M., Ryu J., Kiraly M., Duke K., Reinke V. and Kim S. K. (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in *C. elegans*. *Proc. Natl. Acad. Sci. USA,* **98:** 218-223.

Joen M., Gardner H. F., Miller E. A., Deshler J. and Rougvie A. F. (1999) Similarity of the *C. elegans* developmental timing protein LIN-42 to circadian rhythm proteins. *Science*, **286:** 1141-1146.

Johnstone I. L. and Barry J. D. (1996) Temporal reiteration of a precise gene expression pattern during nematode development. *EMBO J.*, **15:** 3633-3639.

Johnstone I. L. (2000) Cuticle collagen genes. Expression in *Caenorhabditis elegans*. *Trends Genet.*, **16:** 21-7.

Jonassen I., Collins J. F. and Higgins D. G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4:** 1587-1595.

Jubb A. M., Quirke P. and Oates A. J. (2003) DNA methylation, a biomarker for colorectal cancer: implications for screening and pathological utility. *Annals. New York Academy of Science*, **983:** 251-67.

Kamath R. S., Fraser A. G., Dong Y., Poulin G., Durbin R., Gotta M., Kanapin A., Le Bot N., Moreno S., Sohnmann M., Welchman D. P., Zipperfen P. and Ahringer J. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature,* **421:** 231-237.

Kang S. H. and Kramer J. M. (2000) Nidogen is nonessential and not required for normal type IV collagen localization in *Caenorhabditis elegans*. *Mol. Biol. Cell*, **11:** 3911-3923.

Kawano T., Ito Y., Ishiguro M., Takuwa K., Nakajima T. and Kimura Y. (2000) Molecular cloning and characterization of a new insulin/IGF-like peptide of the nematode *Caenorhabditis elegans*. *Biochem. Biophys. Res. Commun.*, **273:** 431-6.

Kelly W. G., Xu S.Q. and Fire A. (1996) Towards an understanding of gene activation and silencing in the nematode germline. *East Coast Worm Meeting*, abstract 67.

Kelly W. G., Xu S.Q., Montgomery M. K. and Fire A. (1997) Distinct Requirements for Somatic and Germline Expression of a Generally Expressed *Caenorhabditis elegans* gene. *Genetics*, **146:** 227-238.

Kelly W. G. and Fire A. (1998) Chromatin silencing and the maintenance of a functional germline in *Caenorhabditis elegans*. *Development*, **125:** 2451-2456.

Kent W. J. and Zahler A. M. (2000) Conservation, regulation, synteny, and introns in a large-scale C.briggsae-*C. elegans* genomic alignment. *Genome Res.*, **10:** 1115-1125.

Kim S. K. and Wadsworth W. G. (2000) Positioning of longitudinal nerves in *C. elegans* by nidogen. *Science*, **288:** 150-154.

Kim S. K., Lund J., Kiraly M., Duke K., Jiang M., Stuart J. M., Eizinger A., Wylie B. N. and Davidson G. S. (2001) A Gene Expression Map for *Caenorhabditis elegans*. *Science*, **293:** 2087-2092.

Kimble L. and Hirsh D. (1979) The postembryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans*. *Devl. Biol.*, **70:** 396-417.

Kirouac M. and Sternberg P. W. (2003) cis-Regulatory control of three cell fate specific genes in vulval organogenesis of *Caenorhabditis elegans*. *Dev. Biol.*, **257:** 85-103.

Klingenhoff A., Frech K., Quandt K., Werner T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15:** 180-186.

Kolchanov N. A., Ponomarenko M. P., Frolov A. S., Ananko E. A., Kolpakov F. A., Ignatieva E. V., Podkolodnaya O. A., Goryachkovskaya T. N., Stepanenko I. L., Merkulova T. I., Babenko V. V., Ponomarenko Y. V., Kochetov A. V., Podkolodny N. L., Vorobiev D. V., Lavryushev S. V., Grigorovich D. A., Kondrakhin Y. V., Milanesi L., Wingender E., Solovyev V., Overton G.C. (1999) Integrated databases and computer systems for studying eukaryotic gene expression. *Bioinformatics*, **15:** 669-686.

Kraemer B., Crittenden S., Gallegos M., Moulder G., Barstead R., Kimble J and Wickens M. (1999) NANOS-3 and FBF proteins physically interact to control the sperm-oocyte switch in *Caenorhabditis elegans*. *Curr. Biol.*, **9:** 1009-18.

Krause M., Fire A., Harrison S. W., Priess J. and Weintraub H. (1990) CeMyoD accumulation defines the body wall muscle cell fate during *C. elegans* embryogenesis. *Cell*, **63:** 907-19.

Lawrence, C. E. and Reilly, A. A. (1990) An expectation maximization (EM) algorithm for the identification and characterisation of common sites in unaligned biopolymer sequences. *PROTEINS: Struct. Funct. Genet.*, **7:** 41-51.

Lawrence C. E., Altschul S. F., Boguski M. S., Liu J. S., Neuwald A. F. and Wootton J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262:** 208-214.

Lee M. H., Park H., Shim G., Lee J. and Koo H. S. (2001) Regulation of gene expression, cellular localization, and in vivo function of *Caenorhabditis elegans* DNA topoisomerase I. *Genes Cells,* **6:** 303-12.

Levitan D., Yu G., St George Hyslop P. and Goutte C. (2001) APH-2/nicastrin functions in LIN-12/Notch signaling in the *Caenorhabditis elegans* somatic gonad. *Dev. Biol.*, **240:** 654-61.

Loots G. G., Locksley R. M., Blankespoor C. M., Wang Z. E., Miller W., Rubin E. M. and Frazer K. A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 15 by cross-species sequence comparisons. *Science,* **288:** 136-140.

Lynch A.S., Briggs D., Hope I.A. (1995) Developmental expression pattern screen for genes predicted in the *C. elegans* genome sequencing project. *Nat. Genet.*, **11:** 309-313.

Maduro M. F., Meneghini M. D., Bowerman B., Broitman-Maduro G and Rothman J. H. (2001) Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in *C. elegans*. *Mol. Cell*, **7:** 475-85.

Maeda I, Kohara Y, Yamamoto M, Sugimoto A. (2001) Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. *Curr. Biol.*, **11:** 171-6.

Makalowski W. and Boguski M. S. (1998) Evolutionary parameters of the transcribed mammalian genome: An analyis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA*, **95:** 9407-9412.

Mcguire A. M. and Church G. M. (2000) Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nucleic Acids Res.*, **28:** 4523-4530.

Mewes H. W., Albermann K., Bahr M., Frishman D., Gleissner A., Hani J., Heumann K., Kleine K., Maierl A., Oliver S. G., Pfeiffer F. and Zollner A. (1997) Overview of the yeast genome. *Nature*, **387:** 7-65. *Erratum in: Nature,* **387:** 737.

Miguel-Aliaga I., Culetto E., Walker D. S., Baylis H. A., Sattelle, D. B. and Davies K. E. (1999) The *Caenorhabditis elegans* orthologue of the human gene responsible for spinal muscular atrophy

is a maternal product critical for germline maturation and embryonic viability. *Hum. Mol. Genet.*, **8:** 2133-2143.

Miller D. M., Stockdale F. E. and Karn J. (1986) Immunological identification of the genes encoding the four myosin heavy chain isoforms of *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U S A*, **83:** 2305-9.

Mir K. U. (2000) The hypothesis is there is no hypothesis. *Trends Genet.*, **16:** 63-64.

Mounsey A. Bauer P and Hope I. A. (2002) Evidence suggesting that a fifth of annotated Caenorhabditis elegans gene may be pseudogenes. *Genome Res.*, **12:** 770-775.

Mount D. W. (2001) Bioinformatics Sequence and Genome Analysis. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York*, 173-204.

Nelson K. F., Albert P. S. and Riddle D. L. (1983) Fine structure of the *C. elegans* secretory-excretory. *J. Ultrastruc. Res.*, **82:** 156-171.

Neuwald A. F. and Green P. (1994) Detecting patterns in protein sequences. *J. Mol. Biol.*, **239:** 689-712.

Newman-Smith E. D., and Rothman J. H. (1998) The maternal-to-zygotic transition in embryonic patterning of *Caenorhabditis elegans*. *Curr. Opin. Genet. Dev.*, **8:** 472-480.

Niehrs C., Pollet N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**(6761)**:** 483-487.

Ogawa H., Harada S., Sassa T., Yamamoto H. and Hosono R. (1998) Functional properties of the unc-64 gene encoding a *Caenorhabditis elegans* syntaxin. *J. Biol. Chem.*, **273:** 2192-8.

Ohler U. and Niemann H. (2001) Identification analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17:** 56-60.

Oishi I., Iwai K., Kagohashi Y., Fujimoto H., Kariya K., Kataoka T., Sawa H., Okano H., Otani H., Yamamura H. And Minami Y. (2001) Critical role of *Caenorhabditis elegans* homologs of Cds1 (Chk2)-related kinases in meiotic recombination. *Mol. Cell. Biol.*, **21:** 1329-35.

Okkema P. G., Harrison SW, Plunger V, Aryana A and Fire A. (1993) Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics*, **135:** 385-404.

Palin K., Ukkonen E., Brazma A. and Vilo J. (2002) Correlating gene promoters and expression in gene disruption experiments. *Bioinformatics*, **18:** S172-S180.

Pearson W. R. and Lipman D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85:** 2444-2448.

Pickert L., Reuter I., Klawonn F. and Wingender E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14:** 244-251.

Pilgrim D. (1998) CeRep25B forms chromosome-specific minisatellite array in *Caenorhabditis elegans*. *Genome Res.*, **8:** 1192-1201.

Plasterk R. H. (1996) Postsequence genetics of *Caenorhabditis elegans*. *Genome Res.*, **6:** 169-75.

Plasterk R. H. A. (1999) Hershey heaven and *Caenorhabditis elegans*. *Nat. Genet.*, **21:** 63-64.

Podbilewicz B. (1996) ADM-1, a protein with metalloprotease-like and disintegrin-like domains, is expressed in syncytial organs, sperm, and sheath cells of sensory organs in *Caenorhabditis elegans*. *Mol. Biol. Cell,* **7:** 1877-1893.

Pollet N., Schmidt H. A., Gawantka V., Vingron M. and Niehrs C. (2000) Axeldb: a *Xenopus laevis* database focusing on gene expression. *Nucleic Acids Res.*, **28:** 139-140.

Poole A. M., Phillips M. J. and Penny D. (2003) Prokaryote and eukaryote evolvability. *Biosystems*, **69:** 163-85.

Praitis V., Casey E., Collar D. and Austin J. (2001) Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics*, **157:** 1217-1226.

Pribnow D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA*, **72:** 784-788.

Puoti A., Pugnale P., Belfiore M., Schlappi A. C. and Saudan Z. (2001) RNA and sex determination in *Caenorhabditis elegans*: Post-transcriptional regulation of the sex-determining gene tra-2 and fem-3 mRNAs in the *Caenorhabditis elegans* hermaphrodite. *EMBO J.*, **2:** 899-904.

Puthalakath H. and Strasser A. (2002) Keeping killers on a tight leash: transcriptional and posttranscriptional control of the pro-apoptotic activity of BH3-only proteins. *Cell Death Differ.*, **9:** 505-512.

Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23:** 4878-4884.

Queen C., Wegman, M. N. and Korn L. J. (1982) Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. *Nucleic Acids Res.*, **10:** 449-456.

Quinlan J. R. (1993) C4.5: Programs for machine learning. *San Mateo, Calif: Morgan Kaufmann.*

Reboul J., Vaglio P., Tzellas N., Thierry-Mieg N., Moore T., Jacjson C., Shin-I T., Kohara Y., Thierry-Mieg D. and Thierry-Mieg J. *et al.* (2001) Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nat. Genet.*, **27:** 332-336.

Reese K. J., Dunn M. A., Waddle J. A. and Seydoux G. (2000) Asymmetric segregation of PIE-1 in C. elegans is mediated by two complementary mechanisms that act through separate PIE-1 protein domains. *Mol. Cell,* **6:** 445-55.

Reinke V., Smith H. E., Nance J., Wang J., Van Doren C. Begley R., Jones S. J. M., Davis E. B., Scherer S., Ward S. and Kim S. K. (2000) A global profile of germ line gene expression in *C. elegans*. *Mol. Cell*, **6:** 605-616.

Renart J., Reiser J. and Stark G. R. (1979) Transfer of proteins from gels to diazobenzyloxymethyl-paper and detection with antisera: a method for studying antibody specificity and antigen structure. *Proc. Natl. Acad. Sci. USA*, **76:** 3116-3120.

Riddle D. L., Blumenthal T., Meyer B. J. and Priess J. R. (1997) Introduction to *Caenorhabditis elegans*. In "*C. elegans*". (Riddle D. L., Meyer B. J. and Priess J. R. Ed) *Cold Spring Harbor Laboratory Press*, 1-22.

Ringwald M., Epping J. T. and Richardson J. E. (2000) GXD: integrated access to gene expression data for the laboratory mouse. *Trends Genet.*, **16:** 188-190.

Ringwald M., Eppig J. T., Begley D. A., Corradi J.P., McCright I. J., Hayamizu T. F., Hill D. P., Kadin J. A. and Richardson J. E. (2001) The Mouse Gene Expression Database (GXD). Nucleic Acids Res., **29:** 98-101.

Roberts L. (1990) The Worm Project, Science. *Science*, **248:** 1310-1313.

Robinson K., McGuire A. M. and Church G. M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284:** 241-254.

Saifee O., Wei L. and Nonet M. L. (1998) The Caenorhabditis elegans unc-64 locus encodes a syntaxin that interacts genetically with synaptobrevin. *Mol. Biol. Cell*, **9:** 1235-52.

Sagot M. F., Viari A. and Soldano H. (1997) Multiple comparison a peptide matching approach. *Theoret. Comput. Sci.*, **180:** 1150137.

Scherf M., Klingenhoff A. and Werner T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297:** 599-606.

Schug J. and Overton G. C. (1997) Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. *Intell. Sys. Mol. Biol.*, **5:** 268-271.

Seydoux G. and Fire A. (1994) Soma-germline asymmetry in the distributions of embryonic RNAs in *Caenorhabditis elegans*. *Development*, **120:** 2823-34.

Seydoux G. and Strome S. (1999) Launching the germline in *Caenorhabditis elegans*: regulation of gene expression in early germ cells. *Development*, **126:** 3275-3283.

Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. and Futcher B. (1998) Comprehensive identification of cell-cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9:** 3273-3297.

Spieth J., Denison K., Kirtland S., Cane J. and Blumenthal T. (1985) The *C. elegans* vitellogenin genes: short sequence repeats in the promoter regions and homology. *Nucleic Acids Res.*, **13(14)**: 5283-5295.

Staden R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12:** 505-519.

Stein L. D. and Thierry-Mieg J. (1998) Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACeDB databases. Genome Res., **8**(12): 1308-1315.

Stein D. L. (1999) Internet access to the *C. elegans* genome. *Trends Genet.*, **15:** 425-427.

Stein L. D. and Thierry-Mieg J . (1999) ACeDB: A Genome Database Management System. *Computing in Science & Engineering*, **May-June:** 44-53.

Stein L., Sternberg P., Durbin R., Thierry-Meig J. and Spieth J. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **29:** 82-86.

Strange K. (2003) From genes to integrative physiology: ion channel and transporter biology in *Caenorhabditis elegans*. *Physiol. Rev.*, **83:** 377-415.

Stormo G. D., Schneider T. D., Gold L. and Ehrenfeucht A. (1982) Use of 'Perceptron' algorithm to distinguish translational initiation sites in *E.coli*. *Nucleic Acids Res.*, **10:** 2997-3012.

Stormo G. D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183:** 211-21.

Stormo G. D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16:** 16-23.

Subramaniam K. and Seydoux G. C. (1999) nos-1 and nos-2, two genes related to Drosophila nanos, regulate primordial germ cell development and survival in *Caenorhabditis elegans*. *Development,* **126:** 4861-4871.

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

Sulston J. E., Albertson D. G. and Thomson J. N. (1980) The *Caenorhabditis elegans* male: postembryonic development of nongonadal structures. *Dev. Biol.*, **78:** 542-576.

Sulston J. E., Schierenberg E., White J. G. and Thomson J. N. (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.*, **100:** 64-119.

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

Sumiyama K., Kim C. B. and Ruddle F. H. (2001) An efficient *cis*-element discovery method using multiple sequence comparisons based on evolutionary relationships. *Genomics*, **71:** 260-2.

Sun Y., Witte D. P., Jin P. and Grabowski G. A. (2003) Analyses of temporal regulatory elements of the prosaposin gene in transgenic mice. *Biochem. J.*, **370:** 557-66.

Surzycki S. and Belknap W. R. (2000) Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes. *Proc. Natl. Acad. Sci. USA*, **97:** 245-249.

Tabara H., Motohashi T. and Kohara, Y. (1996) A multi-well version of *in situ* hybridization on whole mount embryos of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **24:** 2119-2124.

Takemoto T., Sasaki Y., Hamajima N., Goshima Y., Nonaka M. and Kimura H. (2000) Cloning and characterization of the Caenorhabditis elegans CeCRMP/DHP-1 and -2; common ancestors of CRMP and dihydropyrimidinase? *Gene,* **261:** 259-67.

Tcherepanova I., Bhattacharyya L., Rubin C, S. and Freedman J. H. (2000) Aspartic proteases from the nematode *Caenorhabditis elegans*. Structural organization and developmental and cell-specific expression of asp-1. *J. Biol. Chem.*, **275:** 26359-69.

Thatcher J. D., Fernandez A. P., Beaster-Jones L., Haun C. and Okkema P. G. (2001) The *Caenorhabditis elegans peb-1* gene encodes a novel DNA-binding protein involved in morphogenesis of the pharynx, vulva, and hindgut. *Dev. Biol.*, **229:** 480-93.

The *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science,* **282:** 2012-2018.

Thompson J., Gibson T.J., Plewniak F., Jeanmougin F. and Higgins D. G. (1997) CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25:** 4876-4882.

Tuschl T. (2003) RNA sets the standard. *Nature,* **421(16):** 220-221.

Ulyanov A. and Stormo G. D. (1995) Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucleic Acids Res.*, **23:** 1434-1440.

van Helden J., André B. and Collodo-Vides J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281:** 827-842.

Vanet A., Marsen L. Sagot M-F. (1999) Promoter sequences and algorithmical methods for identifying them. *Res. Microbiol.*, **150:** 779-799.

Vilo J., Brazma A., Jonassen I., Robinson A. and Ukkonen E. (2000) Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8:** 384-394.

Vogel B. E. and Hedgecock E. M. (2001) Hemicentin, a conserved extracellular member of the immunoglobulin superfamily, organizes epithelial and other cell attachments into oriented line-shaped junctions. *Development,* **128:** 883-894.

Wang X. and Chamberlin H. M. (2002) Multiple regulatory changes contribute to the evolution of the *Caenorhabditis* lin-48 ovo gene. *Genes Dev.*, **16:** 2345-9.

Wasserman W. W., Palumbo M., Thompson W., Fickett J. W. and Lawrence C. E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26:** 225-228.

Waterston R. and Sulston J. (1995) The Genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA*, **92:** 10836-10840.

Watson A., Smaldon N., Lucke R. and Hawkins T. (1993) The *Caenorhabditis elegans* genome sequencing project: first steps in automation. *Nature*, **362:** 569-570.

Web C. T., Shabalina S. A., Ogurtsov A. Yu. and Kondrashov A. S. (2002) Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae. Nucleic Acids Res.*, **30:** 1233-1239.

White J. G., Southgate E., Thompson J. N. and Brenner S. (1986) The structure of the nervous system of *Caenorhabditis elegans. Philos. R. Trans. Soc. Lon. [Biol.]*, **314:** 1-340.

White J. (1988) The Anatomy. In "The Nematode *Caenorhabditis elegans*". *Cold Spring Harbor Laboratory Press*, 81-122.

Wightman B.C, Ha I. and Ruvkun G. B. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans. Cell*, **75:** 855-862.

Wilkinson H. A. and Greenwald I. (1995) Spatial and temporal patterns of lin-12 expression during *C. elegans* hermaphrodite development. *Genetics,* **141:** 513-26.

Wilson M. A., Hoch R. V., Ashcroft N. R., Kosinski M. E and Golden A. (1999) A *Caenorhabditis elegans* wee1 homolog is expressed in a temporally and spatially restricted pattern during embryonic development. *Biochim. Biophys. Acta.*, **1445:** 99-109.

Wilson R. K. (1999) How the worm was won the *C. elegans* genome sequencing project. *Trends Genet.*, **15:** 51-58.

Wingender E., Kel A. E., Kel O.V., Karas H., Heinemeyer T., Dietze P., Knuppel R., Romaschenko A. G. and Kolchanov N. A. (1997) TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res.*, **25:** 265-8.

Wolfertstetter F, Frech K, Herrmann G, Werner T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12:** 71-80.

Wood W. B., Hecht R., Carr S., Vanderslice R., Wolf N. and Hirsh D. (1980) Parental effects and phenotypic characterization of mutations that affect early development in *Caenorhabditis elegans*. *Dev. Biol.*, **74:** 446-469.

Wootton J. C. and Federhen S. (1993) Statistics of local complexity in amino acid sequences and databases. *Comput. Chem.* **17:** 149-163.

Wootton J. C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem.* **18:** 269-285.

Workman C. T. and Stormo G. D. (2000) ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.*, 467-478.

Young J. M. and Hope I. A. (1993) Molecular markers of differentiation in *Caenorhabditis elegans* obtained by promoter trapping. *Dev. Dyn.*, **196:** 124-132.

Zhang H. and Emmons S. W. (2000) A *C. elegans* mediator protein confers regulatory selectivity on lineage-specific expression of a transcription factor gene. *Genes Dev.*, **14:** 2161-2172.

Zhang L., Wu S. L. and Rubin C. S. (2001) A novel adapter protein employs a phosphotyrosine binding domain and exceptionally basic N-terminal domains to capture and localize an atypical protein kinase C: characterization of *Caenorhabditis elegans* C kinase adapter 1, a protein that avidly binds protein kinase C3. *J. Biol. Chem.*, **276:** 10463-75.

Zhang Y., Ma C., Delohery T., Nasipak B., Foat B. C., Bounoutas A., Bussemaker H. J., Kim S. K. and Chalfie M. (2002) Identification of genes expressed in *C. elegans* touch receptor neurons. *Nature*, **418:** 331-335.

Zhu J. and Zhang M. Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15:** 607-611.

Appendices

List of genes obtained from the published literature and assimilated into WormBase.

Papers	Genes
Baylis HA, Matsuda K, Squire MD, Fleming JT, Harvey RJ, Darlison, Barnard EA, Sattelle DB ACR-3, A Caenorhabditis elegans nicotinic acetylcholine receptor subunit - Molecular cloning and functional expression. **Receptors and Channels 1997 5, 149-158.**	*acr-3*
Seydoux G and Fire A Soma-germline asymmetry in the distribution of embryonic RNAs in *C.elegans.* **Development 1994 120, 2823-2834**	*act-1, CeIF, cey-1, cey-2, crf-2, dpy-30, eft-3, F26E4.8, T02C12.2, hsp-1, C47F8.5, F22F1.1*
Stone S, Shaw JE A *Caenorhabditis elegans* act-4-lacZ fusion - Use as a transformation marker and analysis of tissue-specific expression. **Gene 1993 131, 167-173.**	*act-4*
Korswagen H. C. *et al.* G protein hyperactivation of the *Caenorhabditis elegans* adenylyl cyclase SGS-1 induces neuronal degeneration. **EMBO journal 1998 17, 5059-5065.**	*acy-1*
Podbilewicz B ADM-1, a protein with metalloprotease-like and disintegrin-like domains, is expressed in syncytial organs, sperm, and sheath cells of sensory organs in *Caenorhabditis elegans.* **Molecular Biology of the Cell 1996 7, 1877-1893.**	*adm-1*
Leroux MR, Candido EP Subunit characterization of the *Caenorhabditis* chaperonin containing TCP-1 and expression pattern of the gene encoding CCT-1. **Biochemical and Biophysical Research Communications 1997 241, 687-692.**	*cct-1*
Ashcroft NR, Srayko M, Kosinski ME, Mains PE, Golden A RNA-mediated interference of a cdc25 homolog in *Caenorhabditis elegans* results in defects in the embryonic cortical membrane, meiosis, and mitosis. **Developmental Biology 1999 206, 15-32.**	*cdc-25.1*
Azzaria M, Goszczynski B, Chung MA, Kalb JM, McGhee JD A fork head/HNF-3 homolog expressed in the pharynx and intestine of the *Caenorhabditis elegans* embryo. **Developmental Biology 1996 178, 289-303.**	*fkh-1*
Brunschwig K, Wittmann C, Schnabel R, Burglin TR, Tobler H, Muller F Anterior organization of the *Caenorhabditis elegans* embryo by the labial-like Hox gene ceh-13. **Development 1999 126, 1537-1546.**	*ceh-13*
Okkema PG, Fire A The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. **Development 1994 120, 2175-2186.**	*ceh-22*

Okkema PG, Ha E, Haun C, Chen W, Fire A The *Caenorhabditis elegans* NK-2 homeobox gene ceh-22 activates pharyngeal muscle gene expression in combination with pha-1 and is required for normal pharyngeal development. **Development 1997 124, 3965-3973.**	*ceh-22*
Cassata G, Kagoshima H, Pretot RF, Aspock G, Niklaus G, Burglin TR Rapid expression screening of *Caenorhabditis elegans* homeobox open reading frames using a two-step polymerase chain reaction promotor-GFP reporter construction technique. **Gene 1998 212, 127-135.**	*ceh-38*
Oates AC, Wollberg P, Achen MG, Wilks AF Sampling the genomic pool of protein tyrosine kinase genes using the polymerase chain reaction with genomic DNA. **Biochemical and Biophysical Research Communications 1998 249, 660-667.**	*cehd-5, cehd-7*
Chen W *et al.* A new member of ras superfamily, the rac1 Homologue from *C.elegans.* **JBC 1993 268, 320-324.**	*CErac1*
Morita K, Chow KL, Ueno N Regulation of body length and male tail ray pattern formation of *Caenorhabditis elegans* by a member of the TGF-B family. **Development 1999 126, 1337-1347.**	*cet-1*
Maloof JN, Whangbo J, Harris JM, Jongeward GD, Kenyon C A Wnt signaling pathway controls Hox gene expression and neuroblast migration in *C. elegans.* **Development 1999 126, 37-49.**	*wnt-1, wnt-2*
Hong Y, Roy R, Ambros V Developmental regulation of a cyclin-dependent kinase inhibitor controls postembryonic cell cycle progression in *Caenorhabditis elegans.* **Development 1998 125, 3585-3597.**	*cki-1*
Kramer JM, Cox GN, Hirsh D Expression of the *C. elegans* collagen genes col-1 and col-2 is developmentally regulated. **Journal of Biological Chemistry 1985, 260, 1945-51.**	*col-1, col-2*
Johnstone I L and Barry J D Temporal reiteration of a precise gene expression pattern during nematode development. **The EMBO Journal 1996 15, 3633-3639.**	*col-12, dpy-7*
Levy AD, Kramer JM Identification, sequence and expression patterns of the *Caenorhabditis elegans* col-36 and col-40 collagen-encoding genes. **Gene 1993 137, 281-285.**	*col-3, col-6, col-12, col-13, col-40*
Britton C, McKerrow JH, Johnstone IL Regulation of the *Caenorhabditis elegans* gut cysteine protease gene cpr-1: Requirement for GATA motifs. **Journal of Molecular Biology 1998 283, 15-27.**	*cpr-1*

Larminie CG, Johnstone IL Isolation and characterization of four developmentally regulated cathepsin B-like cysteine protease genes from the nematode *Caenorhabditis elegans*. **DNA and Cell Biology 1996 15, 75-82.**	*cpr-3, cpr-4, cpr-5, cpr-6*
Wakabayashi T, Nakamura N, Sambongi Y, Wada Y, Oka T, Futai M Identification of the copper chaperone, cuc-1, in *Caenorhabditis elegans* - Tissue-specific coexpression with the copper transporting ATPase cua-1. **FEBS Letters 1998 440, 141-146.**	*cuc-1, cua-1*
Page AP, Winter AD Expression pattern and functional significance of a divergent nematode cyclophilin in *Caenorhabditis elegans*. **Molecular and Biochemical Parasitology 1999 99, 301-306.**	*cyp-8*
Ren P, Lim CS, Johnsen R, Albert PS, Pilgrim D, Riddle DL Control of *C. elegans* larval development by neuronal expression of a TGF-Beta homolog. **Science 1996 274, 1389-1391.**	*daf-7*
Suzuki Y, Yandell MD, Roy PJ, Krishna S, Savage-Dunn C, Ross RM, Padgett RW, Wood WB A BMP homolog acts as a dose-dependent regulator of body size and male tail patterning in *Caenorhabditis elegans*. **Development 1999 126, 241-250.**	*dbl-1*
Treinin M and Chalfie M A Mutated Acetylcholine Receptor Subunit Causes Neuronal Degeneration in *C.elegans*. **Neuron 1995 14:871-877.**	*deg-3*
Treinin M, Gillo B, Liebman L, Chalfie M Two functionally dependent acetylcholine subunits are encoded in a single *Caenorhabditis elegans* operon. **Proceedings of the National Academy of Sciences USA 1998 95, 15492-15495.**	*des-2*
Ahringer J Embryonic tissue differentiation in *Caenorhabditis elegans* requires dif-1, a gene homologous to mitochondrial solute carriers. **EMBO Journal 1995 14, 2307-2316.**	*dif-1*
Gilleard JS, Barry JD, Johnstone IL cis regulatory requirements for hypodermal cell-specific expression of the *Caenorhabditis elegans* cuticle collagen gene dpy-7. **Molecular and Cellular Biology 1997 17, 2301-2311.**	*dpy-7*
Clark SG, Shurland DL, Meyerowitz EM, Bargmann CI, van der Bliek AM A dynamin GTPase mutation causes a rapid and reversible temperature-inducible locomotion defect in C. elegans. **Proceedings of the National Academy of Sciences USA 1997 94, 10438-10443.**	*dyn-1*
Ferreira HB, Zhang Y, Zhao C, Emmons SW Patterning of *Caenorhabditis elegans* posterior structures by the Abdominal-B homolog, egl-5. **Developmental Biology 1999 207, 215-228.**	*egl-5*

Jiang LI and Sternberg PW Socket Cells Mediate Spicule Morphogenesis in *Caenorhabditis elegans* Males. **Developmental Biology 1999 211, 88-99.**	*egl-17*
Solari F, Bateman A, Ahringer J The *Caenorhabditis elegans* genes egl-27 and egr-1 are similar to MTA1, a member of a chromatin regulatory complex, and are redundantly required for embryonic patterning. **Development 1999 126, 2483-2494.**	*egl-27*
Baum PD, Guenther C, Frank CA, Pham BV, Garriga The *Caenorhabditis elegans* gene ham-2 links Hox patterning to migration of the HSN neuron. **Genes and Development 1999 13, 472-483G.**	*ham-2, egl-43*
Spieth J, Shim YH, Lea K, Conrad R, Blumenthal T elt-1, an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. **Molecular and Cellular Biology 1991 11, 4651-4659.**	*elt-1*
Fukushige T, Hawkins MG, McGhee JD The GATA-factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. **Developmental Biology 1998 198, 286-302.**	*elt-2*
Gillard JS, Shaffi Y, Barry JD, McGhee JD ELT-3: A *Caenorhabditis elegans* GATA factor expressed in the embryonic epidermis during morphogenesis. **Developmental Biology 1999 208:265-280.**	*elt-3*
Graham PL, Johnson JJ, Wang S, Sibley MH, Gupta MC, Kramer JM Type IV collagen is detectable in most, but not all, basement membranes of *Caenorhabditis elegans* and assembles on tissues that do not suppress it. **Journal of Cell Biology 1997 137, 1171-1183.**	*emb-9, let-2*
Gaudet J, VanderElst I, Spence AM Post-transcriptional regulation of sex determination in *Caenorhabditis elegans* - Widespread expression of the sex-determining gene fem-1 in both sexes. **Molecular Biology of the Cell 1996 7, 1107-1121.**	*fem-1*
Skipper M, Milne CA, Hodgkin J Genetic and molecular analysis of fox-1, a numerator element involved in *Caenorhabditis elegans* primary sex determination. **Genetics 1999 151, 617-631.**	*fox-1*
Wang W, Shakes DC Expression patterns and transcript processing of FTT-1 and FTT-2, two *C. elegans* 14-3-3 homologues. **Journal of Molecular Biology 1997 268, 619-630.**	*fit-1, fit-2*
Ray C, McKerrow JH Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. **Molecular and Biochemical Parasitology 1992 51, 239-250.**	*gcp-1*

Kennedy BP, Aamodt EJ, Allen FL, Chung MA, Heschl MF, McGhee JD The gut esterase gene (ges-1) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. **Journal of Molecular Biology 1993 229, 890-908.**	*ges-1*
Egan CR, Chung MA, Allen FL, Heschl MF, Van Buskirk CL, McGhee JD A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans* ges-1 gene centers on two GATA sequences. **Developmental Biology 1995 170, 397-419.**	*ges-1*
Edgar LG, McGhee JD Embryonic expression of a gut-specific esterase in *Caenorhabditis elegans*. **Developmental Biology 1986 114, 109-118.**	*ges-1*
Jones AR, Francis R, Schedl T GLD-1, a cytoplasmic protein essential for oocyte differentiation, shows stage- and sex-specific expression during *Caenorhabditis elegans* germline development. **Developmental Biology 1996 180, 165-183.**	*gld-1*
Austin J, Kimble J Transcript analysis of glp-1 and lin-12, homologous genes required for cell interactions during development of *C. elegans*. **Cell 1989 58, 565-571.**	*glp-1*
Chen S, Zhou S, Sarkar M, Spence AM, Schachter H Expression of three *Caenorhabditis elegans* N-acetylglucosaminyltransferase I genes during development. **Journal of Biological Chemistry 1999 274, 288-297.**	*gly-12, gly-13, gly-14*
Huang XY, Barrios LA, Vonkhorporn P, Honda S, Albertson DG, Hecht RM Genomic organization of the glyceraldehyde-3-phosphate dehydrogenase gene family of *Caenorhabditis elegans*. **Journal of Molecular Biology 1989 206, 411-424.**	*gpd-1, gpd-2, gpd-3, gpd-4*
Korswagen HC, Park JH, Ohshima Y, Plasterk RH An activating mutation in a *Caenorhabditis elegans* G(s) protein induces neural degeneration. **Genes and Development 1997 11, 1493-1503.**	*gsa-1*
Park JH, Ohshima S, Tani T, Ohshima Y Structure and expression of the gsa-1 gene encoding a G protein alpha(s) subunit in *C. elegans*. **Gene 1997 194, 183-190.**	*gsa-1*
Baum PD, Guenther C, Frank CA, Pham BV, Garriga The *Caenorhabditis elegans* gene ham-2 links Hox patterning to migration of the HSN neuron. **Genes and Development 1999 13, 472-483G.**	*ham-2, egl-43*
Fay DS, Stanley HM, Han M, Wood WB A *Caenorhabditis elegans* homologue of hunchback is required for late stages of development but not early embryonic patterning. **Developmental Biology 1999 205, 240-253.**	*hbl-1*
Harfe BD, Vaz Gomes A, Kenyon C, Liu J, Krause M, Fire A Analysis of a *Caenorhabditis elegans* Twist homolog identifies conserved and divergent aspects of mesodermal patterning. **Genes and Development 1998 12, 2623-2635.**	*hlh-8*

Khan ML, Gogonea CB, Siddiqui ZK, Ali MY, Kikuno R, Nishikawa K, Siddiqui SS Molecular cloning and expression of the *Caenorhabditis elegans* klp-3, an ortholog of C terminus motor kinesins Kar3 and ncd. **Journal of Molecular Biology 1997 270, 627-639.**	*klp-3*
Wissmann A, Ingles J, McGhee JD, Mains PE *Caenorhabditis elegans* LET-502 is related to Rho-binding kinases and human myotonic dystrophy kinase and interacts genetically with a homolog of the regulatory subunit of smooth... **Genes and Development 1997 11:409-422.**	*let-502*
Han M, Sternberg PW Analysis of dominant negative mutations of the *Caenorhabditis elegans* let-60 ras gene. **Genes and Development 1991 5, 2188-2198.**	*let-60*
Dent JA, Han M Post-embryonic expression pattern of *C. elegans* let-60 ras reporter constructs. **Mechanisms of Development 1998 72, 179-182.**	*let-60*
Levitan D, Greenwald I LIN-12 protein expression and localization during vulval development in *C. elegans*. **Development 1998 125, 3101-3109.**	*lin-12*
Ruvkun G, Ambros V, Coulson A, Waterston R, Sulston J, Horvitz HR Molecular genetics of the *Caenorhabditis elegans* heterochronic gene lin-14. **Genetics 1989 121, 501-516.**	*lin-14*
Ruvkun G, Giusto J The *Caenorhabditis elegans* heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. **Nature 1989 338, 313-319.**	*lin-14*
Ruvkun G and Giusto J Dominant gain-of-function mutations that lead to misregulation of the *C. elegans* heterochronic gene lin-14, and the evolutionary implications of dominant mutations in pattern-formation genes. **Development Supplement 1991 1, 47-54.**	*lin-14*
Arasu P, Wightman B, Ruvkun G Temporal regulation of lin-14 by the antagonistic action of two other heterochronic genes, lin-4 and lin-28. **Genes and Development 1991 5, 1825-1833.**	*lin-14*
Miller LM, Gallegos ME, Morisseau BA, Kim SK lin-31, a *Caenorhabditis elegans* HNF-3/fork head transcription factor homolog, specifies three alternative cell fates in vulval development. **Genes and Development 1993 7, 933-947.**	*lin-31*
Joen M, Gardner HF, Miller EA, Deshler J, Rougvie AF Similarity of the *C. elegans* developmental timing protein LIN-42 to circadian rhythm proteins. **Science 1999 286, 1141-1146.**	*lin-42*
Chamberlin HM, Brown KB, Sternberg PW, Thomas JB Characterization of seven genes affecting *Caenorhabditis elegans* hindgut development. **Genetics 1999 153, 731-723.**	*lin-49, lin-59*

Chamberlin HM and Thomas JH The bromodomain protein LIN-49 and trithorax-related protein LIN-59 affect development and gene expression in *C.elegans.* **Development 2000 127, 713-723.**	*lin-49, lin-59*
Salser SJ, Kenyon C Activation of a *C. elegans* Antennapedia homolog in migrating cells controls their direction of migration. **Nature 1992 355, 255-258.**	*mab-5*
Cowing DW, Kenyon C Expression of the homeotic gene mab-5 during *Caenorhabditis elegans* embryogenesis. **Development 1992 116, 481-490.**	*mab-5*
Costa M, Weir M, Coulson A, Sulston J, Kenyon C Posterior pattern formation in *C. elegans* involves position-specific expression of a gene containing a homeobox. **Cell 1988 55, 747-756.**	*mab-5*
Yuan J, Tirabassi RS, Bush AB, Cole MD The *C. elegans* MDL-1 and MXL-1 proteins can functionally substitute for vertebrate MAD and MAX. **Oncogene 1998 17, 1109-1118.**	*mdl-1, mxl-1*
Wissmann A, Ingles J, Mains PE The *Caenorhabditis elegans* mel-11 myosin phosphatase regulatory subunit affects tissue contraction in the somatic gonad and the embryonic epidermis and genetically interacts with the Rac signaling pa **Developmental Biology 1999 209:111-127.**	*mel-11*
Brocks A, Gerrard B, Allikmets R, Dean M, Plasterk RH Homologs of the human multidrug resistance genes MRP and MDR contribute to heavy metal resistance in the soil nematode *Caenorhabditis elegans.* **EMBO Journal 1996 15, 6132-6143.**	*mrp-1*
Hedgecock EM, Herman RK The ncl-1 gene and genetic mosaics of *Caenorhabditis elegans.* **Genetics 1995 141, 989-1006.**	*ncl-1*
Sluder AE, Lindblom T, Ruvkun G The *Caenorhabditis elegans* orphan nuclear hormone receptor gene nhr-2 functions in early embryonic development. **Developmental Biology 1997 184, 303-319.**	*nhr-2*
Collet J, Spike CA, Lundquist EA, Shaw JE, Herman RK Analysis of osm-6, a gene that affects sensory cilium structure and sensory neuron function in *Caenorhabditis elegans.* **Genetics 1998 148, 187-200.**	*osm-6*
Faber PW, Alter JR, MacDonald ME, Hart AC Polyglutamine-mediated dysfunction and apoptotic death of a *Caenorhabditis elegans* sensory neuron. **Proceedings of the National Academy of Sciences USA 1999 96, 179-184.**	*osm-10*
Fitzgerald MC, Schwarzbauer JE Importance of the basement membrane protein SPARC for viability and fertility in *Caenorhabditis elegans.* **Current Biology 1998 8, 1285-1288.**	*ost-1*

Jia Y, Xie G, McDermott JB, Aamodt E The *C. elegans* gene pag-3 is homologous to the zinc finger proto-oncogene gfi-1. **Development 1997 124, 2063-2073.**	*pag-3*
Iino Y, Yamamoto M Expression pattern of the *C. elegans* p21-activated protein kinase, CePAK. **Biochemical and Biophysical Research Communications 1998 245, 177-184.**	*pak-1*
Lincke CR, Broeks A, The I, Plasterk RH, Borst P The expression of 2 P-glycoprotein (pgp) genes in transgenic *Caenorhabditis elegans* is confined to intestinal cells. **EMBO Journal 1993 12, 1615-1620.**	*pgp-1*
Chase D, Serafinas C, Ashcroft N, Kosinski M, Longo D, Ferris DK, Golden A The polo-like kinase PLK-1 is required for nuclear envelope breakdown and the completion of meiosis in *Caenorhabditis elegans.* **Genesis. 2000 26, 26-41.**	*plk-1*
Tabara H, Hill RJ, Mello CC, Priess JR, Kohara Y pos-1 encodes a cytoplasmic zinc-finger protein essential for germline specification in *C. elegans.* **Development 1999 126, 1-11.**	*pos-1*
Zallen JA, Yi BA, Bargmann CI The conserved immunoglobulin superfamily member SAX-3/Robo directs multiple aspects of axon guidance in *C. elegans.* **Cell 1998 92, 217-227.**	*sax-3*
Grant B, Greenwald IS Structure, function, and expression of SEL-1, a negative regulator of LIN-12 and GLP-1 in *C. elegans.* **Development 1997 124:637-644.**	*sel-1*
Korswagen HC, van der Linden AM, Plasterk RH G protein hyperactivation of the *Caenorhabditis elegans* adenylyl cyclase SGS-1 induces neuronal degeneration. **EMBO Journal 1998 17, 5059-5065.**	*sgs-1, acy-2*
Cox GN, Fields C, Kramer JM, Rosenzweig B, Hirsh D Sequence comparisons of developmentally regulated collagen genes of *Caenorhabditis elegans.* **Gene 1989 76, 331-344.**	*skn-1, pal-1, mex-3*
Park YS, Kramer JM The *C. elegans* sqt-1 and rol-6 collagen genes are coordinately expressed during development, but not at all stages that display mutant phenotypes. **Developmental Biology 1994 163, 112-124.**	*sqt-1, rol-6*
Tanaka Y, Ohta A, Matsuo M, Sakamoto H Developmental expression pattern of the *Caenorhabditis elegans* homologue of the Drosophila suppressor of forked gene. **DNA Research 1995 2, 143-146.**	*suf-1*
Fukushige T, Yasuda H, Siddiqui SS Selective expression of the tba-1 alpha tubulin gene in a set of mechanosensory and motor neurons during the development of *Caenorhabditis elegans.* **Biochimica et Biophysica Acta-Gene Structure and Expression 1995 1261, 401-406.**	*tba-1*

Fukushige T, Yasuda H, Siddiqui SS Molecular cloning and developmental expression of the alpha-2 tubulin gene of *Caenorhabditis elegans.* **Journal of Molecular Biology 1993 234, 1290-1300.**	*tba-2*
Fukushige T and Siddiqui SS Effect of the dpy-20 and rol-6 cotransformation markers on a-tubulin gene expression in *C.elegans* transformants **Transgenic Research 1995 4, 332-340.**	*tba-2*
Kagawa H, Sugimoto K, Matsumoto H, Inoue T, Imadzu H, Takuwa K, Sakube Y Genome structure, mapping and expression of the tropomyosin gene tmy-1 of *Caenorhabditis elegans.* **Journal of Molecular Biology 1995 251, 603-613.**	*tmy-1*
Zhen M, Schein JE, Baillie DL, Candido EP An essential ubiquitin-conjugating enzyme with tissue and developmental specificity in the nematode *Caenorhabditis elegans.* **EMBO Journal 1996 15, 3229-3237.**	*ubc-2*
Stringham EG, Jones D, Candido EP Expression of the polyubiquitin-encoding gene (ubq-1) in transgenic *Caenorhabditis elegans.* **Gene 1992 113, 165-173.**	*ubq-1*
Jones D, Stringham EG, Graham RW, Candido EP A portable regulatory element directs specific expression of the *Caenorhabditis elegans* ubiquitin gene ubq-2 in the somatic gonad. **Developmental Biology 1995 171, 60-72.**	*ubq-2*
Ardizzi JP and Epstein HF Immunochemical localization of myosin heavy chain isoforms and paramyosin in developmentally and structurally diverse muscle cell types of the nematode *C.elegans.* **The Journal of Cell Biology 1987 105, 2763-2770.**	*unc-15, unc-54, myo-3*
Honda S, Epstein H Modulation of muscle gene expression in *Caenorhabditis elegans*: Differential levels of transcripts, mRNAs, and polypeptides from thick filament proteins during/ **Proceedings of the National Academy of Sciences USA 1990 87, 876-880**	*unc-15, unc-54, myo-3*
Gengyo-Ando K, Kamiya Y, Yamakawa A, Kodaira K-I, Nishiwaki K, Miwa J, Hori I and Hosono R The *C. elegans* unc-18 gene encodes a protein expressed in motor neurons. **Neuron 1993 11, 703-711**	*unc-18*
Ogura K, Wicky C, Magnenat L, Tobler H, Mori I, Muller F, Ohshima Y *Caenorhabditis elegans* unc-51 gene required for axonal elongation encodes a novel serine/threonine kinase. **Genes and Development 1994 8, 2389-2400.**	*unc-51*
Finney M and Ruvkun G The unc-86 gene product couples cell lineage and cell identity in *C. elegans.* **Cell 1990 63, 895-905.**	*unc-86*

Hobert O, Moerman DG, Clark KA, Beckerle MC, Ruvkun G A conserved LIM protein that affects muscular adherens junction integrity and mechanosensory function in *Caenorhabditis elegans*. **Journal of Cell Biology 1999 144, 45-57.**	*unc-9*
Maduro M, Pilgrim D Identification and cloning of unc-119, a gene expressed in the *Caenorhabditis elegans* nervous system. **Genetics 1995 141, 977-988.**	*unc-119*
Ahringer J Posterior patterning by the *Caenorhabditis elegans* even-skipped homolog vab-7. **Gene & Development 1996 10, 1120-1130.**	*vab-7*
MacMorris M, Broverman S, Greenspoon S, Lea K, Madej C, Blumenthal T, Spieth J Regulation of vitellogenin gene expression in transgenic *Caenorhabditis elegans* - Short sequences required for activation of the vit-2 promoter. **Molecular and Cellular Biology 1992 12, 1652-1662.**	*vit-2*
Wilson MA, Hoch RV, Ashcroft NR, Kosinski ME, Golden A A *Caenorhabditis elegans* wee1 homolog is expressed in a temporally and spatially restricted pattern during embryonic development. **Biochimica et Biophysica Acta - Gene Structure & Expression 1999 1445, 99-109.**	*wee-1*
Shackleford GM, Shivakumar S, Shiue L, Mason J, Kenyon C, Varmus HE Two wnt genes in *Caenorhabditis elegans*. **Oncogene 1993 8, 1857-1864.**	*wnt-1, wnt-2*
Aspock G, Kagoshima H, Niklaus G, and Burglin TR *Caenorhabditis elegans* has scores of hedgehog-related genes: sequence and expression analysis. **Genome Research 1999 9, 909-923.**	*wrt-1,-2,-3,-4,-5,-6,-7,-8 and grd-1, -2, -3, -5, -6, -7, -8, -9*

Motifs identified from the genomic insert approach.

		Positive			Negative			Complexity		> 1kb	>1kb_val	< 1kb	1kb_value	Introns							Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	>3rd intron		exon_value	Pos_Score	Total Score
1	ATCGATCA	10	11	2	2	25	33.387	1.321	9.408	7	245	4	180	0	0	0	0	0	0	0	0	0	38.636	81.43
2	AGTATACT	8	8	2	2	18	23.226	1.255	8.812	2	70	6	270	0	0	0	0	0	0	0	0	0	42.5	74.54
3	AAAAAAATTAA	13	16	3	3	33	45	0.474	1.703	6	210	4	180	2	60	0	0	1	5	0	3	-15	27.5	74.20
4	GTATCAAA	11	13	3	3	26	34.839	1.213	8.426	11	385	0	0	0	0	1	15	0	0	0	1	-5	30.385	73.65
5	TTATTCAAT	11	12	4	4	22	29.032	0.937	5.913	5	175	6	270	0	0	0	0	0	0	0	1	-5	36.667	71.61
6	CGATCAGT	6	6	0	0	18	23.226	1.386	10.003	4	140	2	90	0	0	0	0	0	0	0	0	0	38.333	71.56
7	GCCTAATTT	6	7	0	0	19	24.677	1.273	8.972	5	175	2	90	0	0	0	0	0	0	0	0	0	37.857	71.51
8	TAATATTGA	8	8	1	1	21	27.581	0.965	6.169	6	210	2	90	0	0	0	0	0	0	0	0	0	37.5	71.25
9	TAATAACT	11	14	4	4	24	31.935	0.974	6.254	7	245	5	225	0	0	0	0	0	0	0	2	-10	32.857	71.05
10	AAATTGAATA	11	12	3	3	25	33.387	0.898	5.559	8	280	2	90	0	0	1	15	0	0	0	1	-5	31.667	70.61
11	GAAAACCAC	8	8	0	0	24	31.935	0.937	5.913	5	175	2	90	0	0	0	0	0	0	0	1	-5	32.5	70.35
12	ATAAACTA	10	11	2	2	25	33.387	0.9	5.58	10	350	0	0	0	0	0	0	0	0	0	1	-5	31.364	70.33
13	ATTTGGTG	10	11	2	2	25	33.387	0.974	6.254	6	210	3	135	0	0	0	0	0	0	0	2	-10	30.455	70.10
14	ATAAAATTC	9	10	2	2	22	29.032	0.937	5.913	5	175	4	180	0	0	0	0	0	0	0	1	-5	35	69.95
15	TGATCGAT	10	11	2	2	25	33.387	1.321	9.408	3	105	4	180	0	0	0	0	0	0	0	3	-15	27	69.79
16	CATCATAA	9	10	0	0	28	37.742	1.04	6.849	6	210	1	45	0	0	0	0	1	5	0	2	-10	25	69.59
17	GTCTATA	9	10	2	2	22	29.032	1.277	9.008	4	140	4	180	0	0	0	0	0	0	0	2	-10	31	69.04
18	TCGAATCA	10	11	3	3	22	29.032	1.321	9.408	6	210	3	135	0	0	0	0	0	0	0	2	-10	30.455	68.89
19	ACGCACA	11	12	3	4	23	30.484	1.004	6.526	8	280	2	90	0	0	1	15	0	0	0	1	-5	31.667	68.68
20	AAAAAATATTT	8	11	2	2	21	27.581	0.655	3.353	5	175	4	180	2	60	0	0	0	0	0	0	0	37.727	68.66
21	ATGATGACA	6	7	1	1	16	20.323	1.273	8.972	4	140	3	135	0	0	0	0	0	0	0	0	0	39.286	68.58
22	AAATTTTTAG	10	10	2	2	24	31.935	0.943	5.972	5	175	1	45	3	90	0	0	0	0	0	1	-5	30.5	68.41
23	ATTTGCAAT	9	9	1	1	24	31.935	1.215	8.443	4	140	2	90	1	30	0	0	0	0	0	2	-10	27.778	68.16
24	GACTGTA	9	10	4	4	16	20.323	1.352	9.689	7	245	3	135	0	0	0	0	0	0	0	0	0	38	68.01
25	AACAGTAA	11	11	4	4	21	27.581	1.074	7.157	8	280	2	90	0	0	0	0	0	0	0	1	-5	33.182	67.92
26	TTAGTGC	6	7	1	1	16	20.323	1.277	9.008	3	105	3	135	1	30	0	0	0	0	0	0	0	38.571	67.90
27	AATTTTCAGAA	6	6	0	0	18	23.226	1.162	7.964	5	175	1	45	0	0	0	0	0	0	0	0	0	36.667	67.86
28	AATATTTTTTC	7	7	0	0	21	27.581	0.86	5.214	4	140	2	90	0	0	1	15	0	0	0	0	0	35	67.79
29	AATAATTAT	13	13	3	3	30	40.645	0.687	3.639	9	315	0	0	0	0	0	0	1	5	0	3	-15	23.462	67.75
30	GATCGATA	8	8	2	2	18	23.226	1.321	9.408	3	105	4	180	0	0	0	0	0	0	0	1	-5	35	67.63
31	TCACATTGA	6	7	0	0	19	24.677	1.311	9.316	3	105	3	135	0	0	0	0	0	0	0	1	-5	33.571	67.56
32	TGCTGATGC	6	6	0	0	18	23.226	1.311	9.316	6	210	0	0	0	0	0	0	0	0	0	0	0	35	67.54
33	GAACTTTA	9	10	3	3	19	24.677	1.255	8.812	6	210	3	135	0	0	0	0	0	0	0	1	-5	34	67.49
34	CAAACAAC	8	9	1	1	22	29.032	0.662	3.408	9	315	0	0	0	0	0	0	0	0	0	0	0	35	67.44
35	ACATAGAA	7	8	0	0	22	29.032	1.074	7.157	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	67.44
36	TGCACCAA	8	8	0	0	24	31.935	1.255	8.812	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	67.00
37	CAACACAAA	9	9	2	2	21	27.581	0.637	3.18	8	280	1	45	0	0	0	0	0	0	0	0	0	36.111	66.87
38	TATTCTTTT	8	8	0	0	24	31.935	0.684	3.61	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	66.80
39	CTTATATT	10	11	2	2	25	33.387	0.9	5.58	9	315	0	0	0	0	0	0	0	0	0	2	-10	27.727	66.69
40	GTTATTCAA	9	9	3	3	18	23.226	1.215	8.443	9	315	0	0	0	0	0	0	0	0	0	0	0	35	66.67
41	AAAATTGAATA	9	9	1	1	24	31.935	0.86	5.214	6	210	1	45	0	0	1	15	0	0	0	1	-5	29.444	66.59
42	AAAAATCGGA	6	8	1	1	17	21.774	1.089	7.297	6	210	2	90	0	0	0	0	0	0	0	0	0	37.5	66.57
43	AAAGATAT	12	13	5	5	22	29.032	0.9	5.58	7	245	4	180	0	0	0	0	0	0	0	2	-10	31.923	66.54
44	AAGATATC	8	11	1	1	24	31.935	1.213	8.426	7	245	1	45	0	0	0	0	1	5	0	2	-10	25.909	66.27
45	TAGTTTAT	10	11	2	2	25	33.387	0.9	5.58	8	280	0	0	1	30	0	0	0	0	0	2	-10	27.273	66.24
46	ATGTCTGA	6	7	1	1	16	20.323	1.321	9.408	6	210	1	45	0	0	0	0	0	0	0	0	0	36.429	66.16

		Positive			Negative			Complexity		> 1kb	>1kb_val	< 1kb	1kb_value	Introns							Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	>3rd intron		exon_value	Pos_Score	Total Score
47	AAAAAATTTTTT	6	14	0	0	26	34.839	0.693	3.696	2	70	6	270	1	30	2	30	0	0	0	3	-15	27.5	66.03
48	AGATCGAT	7	8	0	0	22	29.032	1.321	9.408	4	140	2	90	0	0	0	0	0	0	0	2	-10	27.5	65.94
49	CATGTTTTT	9	11	3	3	20	26.129	1.003	6.512	10	350	0	0	0	0	1	15	0	0	0	0	0	33.182	65.82
50	ATATATATA	8	11	2	2	21	27.581	0.687	3.639	4	140	3	135	3	90	1	15	0	0	0	0	0	34.545	65.77
51	CTAGTTTTT	8	10	2	2	20	26.129	1.003	6.512	7	245	2	90	0	0	0	0	0	0	0	1	-5	33	65.64
52	ATTATGAG	7	7	1	1	18	23.226	1.082	7.236	7	245	0	0	0	0	0	0	0	0	0	0	0	35	65.46
53	TTGATTCG	9	9	2	2	21	27.581	1.213	8.426	4	140	3	135	0	0	0	0	0	0	0	2	-10	29.444	65.45
54	AAATTGCCA	9	11	2	2	23	30.484	1.273	8.972	6	210	2	90	0	0	0	0	0	0	0	3	-15	25.909	65.37
55	ATATAGTA	6	7	1	1	16	20.323	0.974	6.254	3	105	3	135	1	30	0	0	0	0	0	0	0	38.571	65.15
56	TAAAGCC	10	12	5	5	17	21.774	1.277	9.008	8	280	3	135	0	0	0	0	0	0	0	1	-5	34.167	64.95
57	TTTATGATG	7	7	0	0	21	27.581	0.995	6.442	5	175	1	45	0	0	0	0	0	0	0	1	-5	30.714	64.74
58	AAAAATCGG	9	14	5	5	17	21.774	1.149	7.844	9	315	4	180	0	0	0	0	0	0	0	1	-5	35	64.62
59	GAAAACAGT	7	8	2	2	16	20.323	1.149	7.844	7	245	1	45	0	0	0	0	0	0	0	0	0	36.25	64.42
60	AAAATGAGC	7	8	2	2	16	20.323	1.149	7.844	7	245	1	45	0	0	0	0	0	0	0	0	0	36.25	64.42
61	ACATCAAG	7	8	1	1	19	24.677	1.213	8.426	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	64.35
62	TAAAATTACA	5	6	0	0	16	20.323	0.898	5.559	4	140	2	90	0	0	0	0	0	0	0	0	0	38.333	64.22
63	GTAAGTTA	10	11	4	4	19	24.677	1.082	7.236	9	315	1	45	0	0	0	0	0	0	0	1	-5	32.273	64.19
64	TCATTTTAA	11	11	5	5	18	23.226	0.937	5.913	6	210	4	180	0	0	0	0	0	0	0	1	-5	35	64.14
65	TCCTAATA	6	7	1	1	16	20.323	1.082	7.236	6	210	1	45	0	0	0	0	0	0	0	0	0	36.429	63.99
66	ATAAATTAT	11	13	4	5	21	27.581	0.687	3.639	6	210	5	225	0	0	0	0	0	0	0	2	-10	32.692	63.91
67	ATGCACC	11	12	4	4	22	29.032	1.277	9.008	5	175	3	135	0	0	1	15	0	0	0	3	-15	25.833	63.87
68	TCAAAATTTC	11	12	3	3	25	33.387	1.055	6.987	6	210	2	90	0	0	0	0	0	0	0	4	-20	23.333	63.71
69	TTTAATTTTTT	12	14	5	5	23	30.484	0.474	1.703	11	385	1	45	0	0	1	15	0	0	0	1	-5	31.429	63.62
70	AGAAAAATTT	7	8	2	2	16	20.323	0.898	5.559	3	105	4	180	0	0	1	15	0	0	0	0	0	37.5	63.38
71	ATCAAATGT	7	9	3	3	14	17.419	1.215	8.443	7	245	2	90	0	0	0	0	0	0	0	0	0	37.222	63.08
72	AATATAAG	10	11	2	2	25	33.387	0.9	5.58	8	280	0	0	0	0	0	0	0	0	0	3	-15	24.091	63.06
73	CAAGAAAAC	8	8	3	3	15	18.871	0.849	5.111	5	175	3	135	0	0	0	0	0	0	0	0	0	38.75	62.73
74	AAAATCAGA	8	11	2	2	21	27.581	1.003	6.512	8	280	1	45	0	0	0	0	0	0	0	2	-10	28.636	62.73
75	GATGACAA	10	10	4	4	18	23.226	1.213	8.426	4	140	4	180	0	0	0	0	0	0	0	2	-10	31	62.65
76	AAAATCGGA	8	10	3	3	17	21.774	1.149	7.844	7	245	2	90	0	0	0	0	0	0	0	1	-5	33	62.62
77	TTGGGAAAA	9	12	4	5	16	20.323	1.061	7.041	7	245	4	180	0	0	0	0	0	0	0	1	-5	35	62.36
78	AATAGTTTG	5	7	1	1	14	17.419	1.061	7.041	5	175	2	90	0	0	0	0	0	0	0	0	0	37.857	62.32
79	TGTTTTATA	9	10	2	2	22	29.032	0.849	5.111	7	245	1	45	0	0	0	0	0	0	0	2	-10	28	62.14
80	TTCGTGCA	6	7	1	1	16	20.323	1.321	9.408	4	140	2	90	0	0	0	0	0	0	0	1	-5	32.143	61.87
81	TGAATACAA	7	7	0	0	21	27.581	1.149	7.844	3	105	2	90	0	0	0	0	0	0	0	2	-10	26.429	61.85
82	GAATAATTA	7	8	1	1	19	24.677	0.937	5.913	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	61.84
83	AGGGAAAAA	7	7	0	0	21	27.581	0.637	3.18	5	175	1	45	0	0	0	0	0	0	0	1	-5	30.714	61.48
84	ATGATAACA	6	8	0	0	20	26.129	1.149	7.844	4	140	2	90	0	0	0	0	0	0	0	2	-10	27.5	61.47
85	TTTTTTTGGAAA	5	7	0	0	17	21.774	0.96	6.12	3	105	3	135	0	0	0	0	0	0	0	1	-5	33.571	61.47
86	ACAAACGG	8	8	2	2	18	23.226	1.04	6.849	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	61.32
87	GCACTAAA	5	6	0	0	16	20.323	1.213	8.426	1	35	3	135	1	30	0	0	0	0	0	1	-5	32.5	61.25
88	AAAATTAATTT	8	11	2	2	21	27.581	0.689	3.658	0	0	7	315	1	30	0	0	0	0	0	3	-15	30	61.24
89	GATTTTTTCA	10	12	5	5	17	21.774	1.089	7.297	6	210	3	135	1	30	1	15	0	0	0	1	-5	32.083	61.15
90	AAATTTCCAAA	6	7	1	1	16	20.323	0.995	6.442	7	245	0	0	1	30	0	0	0	0	0	0	0	34.375	61.14
91	TTTTTGTTGT	7	9	2	2	17	21.774	0.5	1.942	7	245	2	90	0	0	0	0	0	0	0	0	0	37.222	60.94
92	AAATTTCCAA	8	11	3	3	18	23.226	1.03	6.758	9	315	0	0	1	30	0	0	0	0	0	1	-5	30.909	60.89
93	CCTATTAT	7	7	3	3	12	14.516	1.04	6.849	4	140	3	135	0	0	0	0	0	0	0	0	0	39.286	60.65
94	TCTTTTCCA	7	8	1	1	19	24.677	0.937	5.913	2	70	4	180	0	0	0	0	0	0	0	2	-10	30	60.59
95	CACGAAAA	12	20	7	7	23	30.484	0.9	5.58	11	385	3	135	0	0	0	0	0	0	0	6	-30	24.5	60.56

		Positive			Negative			Complexity		> 1kb	>1kb_val	< 1kb	1kb_value	Introns							Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	>3rd intron		exon_value	Pos_Score	Total Score
96	TTCAAAATTTC	9	10	3	3	19	24.677	1.036	6.817	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	60.49
97	CGGAGAAAA	7	9	2	2	17	21.774	0.937	5.913	6	210	2	90	0	0	0	0	0	0	0	1	-5	32.778	60.47
98	CAAAAATCA	11	11	4	4	21	27.581	0.849	5.111	6	210	2	90	0	0	1	15	0	0	0	2	-10	27.727	60.42
99	ACCTTGAA	8	9	4	4	13	15.968	1.321	9.408	9	315	0	0	0	0	0	0	0	0	0	0	0	35	60.38
100	AATTGAATAA	9	9	2	2	21	27.581	0.898	5.559	3	105	3	135	0	0	1	15	0	0	0	2	-10	27.222	60.36
101	CGTCAAAA	8	11	3	3	18	23.226	1.213	8.426	8	280	1	45	0	0	0	0	0	0	0	2	-10	28.636	60.29
102	GACGAGAC	6	8	1	1	17	21.774	1.082	7.236	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	60.26
103	AAATTTTAAAA	12	15	6	6	21	27.581	0.655	3.353	9	315	3	135	0	0	0	0	0	0	0	3	-15	29	59.93
104	AAAATTTTTA	10	13	3	3	24	31.935	0.693	3.696	9	315	0	0	0	0	1	15	0	0	0	3	-15	24.231	59.86
105	TATTTAAATA	6	8	0	0	20	26.129	0.693	3.696	7	245	0	0	0	0	0	0	0	0	0	1	-5	30	59.82
106	TTTGTCGT	9	9	3	4	16	20.323	0.9	5.58	5	175	3	135	0	0	0	0	0	0	0	1	-5	33.889	59.79
107	AATACCAAA	7	9	3	3	14	17.419	0.849	5.111	7	245	2	90	0	0	0	0	0	0	0	0	0	37.222	59.75
108	AATTTTAACT	9	9	2	2	21	27.581	0.943	5.972	7	245	0	0	0	0	0	0	0	0	0	2	-10	26.111	59.66
109	ATATCAGA	12	13	6	6	19	24.677	1.213	8.426	9	315	1	45	0	0	0	0	0	0	0	3	-15	26.538	59.64
110	GAAAATAGA	10	10	4	4	18	23.226	0.849	5.111	4	140	4	180	0	0	0	0	0	0	0	2	-10	31	59.34
111	AGTGTTTTTT	6	7	1	1	16	20.323	0.802	4.684	1	35	4	180	1	30	0	0	0	0	0	1	-5	34.286	59.29
112	ATGATCAT	8	10	4	4	14	17.419	1.255	8.812	7	245	2	90	0	0	0	0	0	0	0	1	-5	33	59.23
113	ATTGTTATC	9	9	2	2	21	27.581	1.149	7.844	2	70	3	135	0	0	0	0	0	0	0	3	-15	23.75	59.17
114	GATACTG	10	12	4	4	20	26.129	1.352	9.689	6	210	2	90	0	0	0	0	0	0	0	4	-20	23.333	59.15
115	CACCAAATA	6	6	0	0	18	23.226	0.937	5.913	4	140	1	45	0	0	0	0	0	0	0	1	-5	30	59.14
116	AGCTGCG	8	8	3	3	15	18.871	1.277	9.008	6	210	1	45	0	0	0	0	0	0	0	1	-5	31.25	59.13
117	ATCGGCG	8	11	3	3	18	23.226	1.277	9.008	5	175	3	135	0	0	0	0	0	0	0	3	-15	26.818	59.05
118	TAAAACATT	10	13	6	6	15	18.871	0.937	5.913	9	315	3	135	0	0	0	0	0	0	0	1	-5	34.231	59.02
119	TGAAAATGA	11	14	6	6	18	23.226	0.995	6.442	7	245	4	180	0	0	0	0	0	0	0	3	-15	29.286	58.95
120	TGTAACAAA	7	7	1	1	18	23.226	1.149	7.844	4	140	1	45	0	0	1	15	0	0	0	1	-5	27.857	58.93
121	AGTGTTTTTTT	5	6	0	0	16	20.323	0.76	4.3	0	0	4	180	1	30	0	0	0	0	0	1	-5	34.167	58.79
122	AAAAGTAAT	9	10	3	3	19	24.677	0.849	5.111	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	58.79
123	TGAGTAAT	9	11	5	5	14	17.419	1.082	7.236	7	245	3	135	0	0	0	0	0	0	0	1	-5	34.091	58.75
124	GCAGCTG	12	15	6	7	19	24.677	1.277	9.008	10	350	1	45	0	0	0	0	0	0	0	4	-20	25	58.69
125	TCATACTT	6	8	1	1	17	21.774	1.04	6.849	7	245	0	0	0	0	0	0	0	0	0	1	-5	30	58.62
126	ATTTGGAAAA	7	9	3	3	14	17.419	1.03	6.758	5	175	2	90	1	30	1	15	0	0	0	0	0	34.444	58.62
127	ATTTTAGAAA	7	7	1	1	18	23.226	0.943	5.972	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	58.48
128	GTTACGGT	6	7	1	1	16	20.323	1.255	8.812	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	58.42
129	CATACGG	8	9	2	2	19	24.677	1.352	9.689	5	175	1	45	0	0	0	0	1	5	0	2	-10	23.889	58.25
130	TTTTTTTTCCA	7	8	3	2	15	18.871	0.76	4.3	5	175	1	45	2	60	0	0	0	0	0	0	0	35	58.17
131	ATCATCTA	8	8	4	4	12	14.516	1.082	7.236	7	245	1	45	0	0	0	0	0	0	0	0	0	36.25	58.00
132	AATTATTTAA	6	8	1	1	17	21.774	0.693	3.696	5	175	2	90	0	0	0	0	0	0	0	1	-5	32.5	57.97
133	TAGATTTAT	8	8	2	2	18	23.226	0.937	5.913	3	105	3	135	0	0	0	0	0	0	0	2	-10	28.75	57.89
134	TTCGACGA	10	11	4	4	19	24.677	1.386	10.003	6	210	1	45	0	0	1	15	0	0	0	3	-15	23.182	57.86
135	AACTCGGA	8	9	2	3	17	21.774	1.321	9.408	5	175	1	45	1	30	0	0	0	0	0	2	-10	26.667	57.85
136	CAAAATTTCA	6	7	1	1	16	20.323	1.03	6.758	5	175	1	45	0	0	0	0	0	0	0	1	-5	30.714	57.79
137	TATGGAAAA	7	9	2	2	17	21.774	0.995	6.442	4	140	3	135	0	0	0	0	0	0	0	2	-10	29.444	57.66
138	AAAAAATCGT	8	11	2	2	21	27.581	1.089	7.297	3	105	3	135	1	30	0	0	0	0	0	4	-20	22.727	57.60
139	CACCAGGA	7	10	2	2	18	23.226	1.082	7.236	8	280	0	0	0	0	0	0	0	0	0	2	-10	27	57.46
140	ATCAAAACA	8	9	2	3	17	21.774	0.849	5.111	3	105	4	180	0	0	0	0	0	0	0	2	-10	30.556	57.44
141	GTCTTGCA	6	8	1	1	17	21.774	1.321	9.408	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	57.43
142	ACAACTAC	7	7	1	1	18	23.226	0.974	6.254	4	140	1	45	0	0	1	15	0	0	0	1	-5	27.857	57.34
143	AAAAAATTTTT	8	16	3	4	21	27.581	0.689	3.658	3	105	6	270	1	30	2	30	0	0	0	4	-20	25.938	57.18
144	AACCAAATTT	7	7	1	1	18	23.226	1.03	6.758	1	35	3	135	1	30	0	0	0	0	0	2	-10	27.143	57.13

		Positive			Negative			Complexity		> 1kb	>1kb_val	< 1kb	1kb_value	Introns						>3rd	Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	intron		exon_value	Pos_Score	**Total Score**
145	CAAAAAAGA	10	12	5	5	17	21.774	0.684	3.61	6	210	4	180	0	0	0	0	0	0	0	2	-10	31.667	57.05
146	AAAATTCTA	8	10	2	2	20	26.129	0.937	5.913	5	175	2	90	0	0	0	0	0	0	0	3	-15	25	57.04
147	AAAACGAAAA	11	13	5	5	20	26.129	0.639	3.203	6	210	3	135	1	30	0	0	0	0	0	3	-15	27.692	57.02
148	CTACATAC	6	7	0	0	19	24.677	1.082	7.236	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	56.91
149	CACAAAAAAC	6	7	0	0	19	24.677	0.611	2.947	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	56.91
150	ATTTGTCG	8	10	4	4	14	17.419	1.213	8.426	9	315	0	0	0	0	0	0	0	0	0	1	-5	31	56.85
151	ACAAATGG	9	11	3	3	20	26.129	1.213	8.426	6	210	1	45	0	0	0	0	1	5	0	3	-15	22.273	56.83
152	GATAGAC	8	10	3	3	17	21.774	1.277	9.008	6	210	1	45	0	0	1	15	0	0	0	2	-10	26	56.78
153	GTTTTTCTC	7	8	3	3	13	15.968	0.849	5.111	6	210	1	45	1	30	0	0	0	0	0	0	0	35.625	56.70
154	TTGGGAAAAA	6	8	2	2	14	17.419	1.03	6.758	5	175	2	90	0	0	0	0	0	0	0	1	-5	32.5	56.68
155	TCATGATT	11	14	6	6	18	23.226	1.213	8.426	5	175	3	135	2	60	0	0	0	0	0	4	-20	25	56.65
156	AATAAATTAAT	6	6	0	0	18	23.226	0.655	3.353	4	140	1	45	0	0	0	0	0	0	0	1	-5	30	56.58
157	TGTTAACT	8	11	5	5	12	14.516	1.213	8.426	6	210	3	135	1	30	0	0	0	0	0	1	-5	33.636	56.58
158	TTTATGTG	8	9	3	3	16	20.323	0.9	5.58	8	280	0	0	0	0	0	0	0	0	0	1	-5	30.556	56.46
159	TAAAAAAAATT	10	10	4	4	18	23.226	0.586	2.72	8	280	0	0	1	30	0	0	0	0	0	1	-5	30.5	56.45
160	TGATCAAT	9	10	6	6	10	11.613	1.255	8.812	4	140	5	225	0	0	0	0	0	0	0	1	-5	36	56.43
161	TTTCAGTGA	7	11	2	2	19	24.677	1.273	8.972	4	140	1	45	1	30	3	45	0	0	0	2	-10	22.727	56.38
162	ATAAATGTGA	5	6	0	0	16	20.323	1.03	6.758	0	0	2	90	3	90	0	0	0	0	0	1	-5	29.167	56.25
163	TAAATTAATTT	7	7	1	1	18	23.226	0.689	3.658	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	56.17
164	TTAATTTTTTT	7	8	1	1	19	24.677	0.451	1.488	4	140	2	90	0	0	1	15	0	0	0	1	-5	30	56.17
165	GTAATAAT	10	11	5	5	16	20.323	0.974	6.254	7	245	2	90	0	0	0	0	0	0	0	2	-10	29.545	56.12
166	TCACACAT	6	6	1	1	15	18.871	1.082	7.236	4	140	1	45	0	0	0	0	0	0	0	1	-5	30	56.11
167	TTTTTTAATTTT	8	9	2	3	17	21.774	0.451	1.488	6	210	2	90	0	0	0	0	0	0	0	1	-5	32.778	56.04
168	GGTGCTA	9	9	3	4	16	20.323	1.277	9.008	5	175	1	45	1	30	0	0	0	0	0	2	-10	26.667	56.00
169	AAAGTTCG	8	10	4	4	14	17.419	1.321	9.408	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	55.83
170	TTGTACAA	5	6	1	1	13	15.968	1.255	8.812	2	70	2	90	0	0	0	0	0	0	0	1	-5	31	55.78
171	CACTGAAAA	7	9	2	2	17	21.774	1.149	7.844	7	245	0	0	0	0	0	0	0	0	0	2	-10	26.111	55.73
172	TGATGACAA	7	7	2	2	15	18.871	1.273	8.972	2	70	3	135	0	0	0	0	0	0	0	2	-10	27.857	55.70
173	AATTTTGTAA	7	8	1	1	19	24.677	0.943	5.972	6	210	0	0	0	0	0	0	0	0	0	2	-10	25	55.65
174	CACACGA	11	13	6	6	17	21.774	1.004	6.526	8	280	2	90	0	0	0	0	0	0	0	3	-15	27.308	55.61
175	TTTTAGAAAT	6	7	1	1	16	20.323	0.943	5.972	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	55.58
176	AACAGGAA	9	10	5	5	13	15.968	0.9	5.58	6	210	3	135	0	0	0	0	0	0	0	1	-5	34	55.55
177	CACGAAAAA	6	11	2	2	17	21.774	0.849	5.111	8	280	1	45	0	0	0	0	0	0	0	2	-10	28.636	55.52
178	TATCACCA	7	7	2	2	15	18.871	1.082	7.236	6	210	0	0	0	0	0	0	0	0	0	1	-5	29.286	55.39
179	ATCTGTAA	7	8	2	2	16	20.323	1.255	8.812	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	55.38
180	TAAATTAATTTA	6	6	0	0	18	23.226	0.693	3.696	5	175	0	0	0	0	0	0	0	0	0	1	-5	28.333	55.25
181	TAGATTGT	6	9	2	2	15	18.871	1.04	6.849	4	140	3	135	0	0	0	0	0	0	0	2	-10	29.444	55.16
182	GAAAAAATCG	7	8	3	3	13	15.968	1.089	7.297	4	140	2	90	1	30	0	0	0	0	0	1	-5	31.875	55.14
183	AAAAATAGTT	8	8	2	2	18	23.226	0.898	5.559	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	55.03
184	TTAACAAAT	7	9	3	3	14	17.419	0.937	5.913	7	245	1	45	0	0	0	0	0	0	0	1	-5	31.667	55.00
185	ATATCAAG	7	8	2	2	16	20.323	1.213	8.426	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	55.00
186	TTTAGTCA	8	9	4	4	13	15.968	1.213	8.426	3	105	4	180	0	0	0	0	0	0	0	2	-10	30.556	54.95
187	TTTTTTGGAAA	7	11	3	3	16	20.323	0.995	6.442	4	140	3	135	1	30	1	15	0	0	0	2	-10	28.182	54.95
188	GCAAAAATT	9	11	4	4	17	21.774	1.149	7.844	7	245	1	45	0	0	0	0	0	0	0	3	-15	25	54.62
189	TCTTGTTCA	9	10	3	3	19	24.677	1.149	7.844	3	105	3	135	0	0	0	0	0	0	0	4	-20	22	54.52
190	AAAATTAATTTT	6	8	0	0	20	26.129	0.693	3.696	0	0	4	180	1	30	0	0	0	0	0	3	-15	24.375	54.20
191	AAAATGACA	9	11	4	4	17	21.774	1.003	6.512	6	210	2	90	0	0	0	0	0	0	0	3	-15	25.909	54.20
192	AAAATTTCAAA	8	13	3	5	16	20.323	0.86	5.214	8	280	1	45	0	0	0	0	0	0	0	2	-10	28.636	54.17
193	AAAATCGTA	9	10	4	4	16	20.323	1.149	7.844	4	140	3	135	0	0	0	0	0	0	0	3	-15	26	54.17

		Positive		Negative				Complexity		> 1kb	>1kb val	< 1kb	1kb value	Introns							Exon			
#	Motif	Frag Freq	Tot Freq	Frag Freq	Tot Freq	Freq Val	Freq Score	Value	Score					1st intron	Intron value	2nt intron	Intron val	3rd intron	Intron va	>3rd intron		exon value	Pos Score	Total Score
194	TAGCTAAA	6	6	1	1	15	18.871	1.213	8.426	1	35	3	135	0	0	0	0	0	0	0	2	-10	26.667	53.96
195	AAATCCTCA	7	7	0	0	21	27.581	1.061	7.041	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	53.91
196	AAATGCATC	6	8	2	2	14	17.419	1.273	8.972	6	210	0	0	0	0	1	15	0	0	0	1	-5	27.5	53.89
197	AATACCAA	9	13	7	7	10	11.613	0.9	5.58	10	350	2	90	0	0	0	0	0	0	0	0	0	36.667	53.86
198	TAGTGTAA	4	6	0	0	14	17.419	1.082	7.236	3	105	1	45	1	30	0	0	0	0	0	1	-5	29.167	53.82
199	TAAATGTGA	6	8	1	1	17	21.774	1.061	7.041	0	0	2	90	4	120	0	0	0	0	0	2	-10	25	53.82
200	ATACAGTA	7	8	2	2	16	20.323	1.213	8.426	6	210	0	0	0	0	0	0	0	0	0	2	-10	25	53.75
201	GCAGCTGC	6	6	0	0	18	23.226	1.255	8.812	4	140	0	0	0	0	0	0	0	0	0	2	-10	21.667	53.70
202	ATACGGTAG	5	6	1	1	13	15.968	1.311	9.316	5	175	0	0	0	0	0	0	0	0	0	1	-5	28.333	53.62
203	ATGTATGTA	6	6	0	0	18	23.226	1.061	7.041	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	53.60
204	AAACATTAT	8	9	4	4	13	15.968	0.937	5.913	7	245	1	45	0	0	0	0	0	0	0	1	-5	31.667	53.55
205	ATTTTTTGCA	8	10	4	4	14	17.419	1.089	7.297	5	175	2	90	2	60	0	0	0	0	0	2	-10	28.636	53.35
206	ATTTCAAAAG	6	7	1	1	16	20.323	1.168	8.019	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	53.34
207	TCAATGTA	5	8	2	2	12	14.516	1.255	8.812	7	245	0	0	0	0	0	0	0	0	0	1	-5	30	53.33
208	AAAATTTCCA	5	10	3	3	11	13.065	1.03	6.758	7	245	0	0	3	90	0	0	0	0	0	0	0	33.5	53.32
209	CGATTTTTCA	6	10	2	2	16	20.323	1.221	8.495	4	140	2	90	1	30	0	0	0	0	0	3	-15	24.5	53.32
210	CGGAGAAA	9	11	5	5	14	17.419	0.974	6.254	7	245	2	90	0	0	0	0	0	0	0	2	-10	29.545	53.22
211	AACTACAT	9	10	4	4	16	20.323	1.04	6.849	4	140	3	135	0	0	0	0	0	0	0	3	-15	26	53.17
212	TAGTAATG	7	8	1	1	19	24.677	1.082	7.236	4	140	1	45	0	0	0	0	0	0	0	3	-15	21.25	53.16
213	AATCCTCA	7	8	1	1	19	24.677	1.082	7.236	4	140	1	45	0	0	0	0	0	0	0	3	-15	21.25	53.16
214	CAAACAATT	8	10	4	4	14	17.419	0.995	6.442	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	52.86
215	GATTTGCCGG	2	7	1	1	8	8.71	1.28	9.034	7	245	0	0	0	0	0	0	0	0	0	0	0	35	52.74
216	CATTGTAA	7	7	2	2	15	18.871	1.255	8.812	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	52.68
217	GTAGATTG	6	7	1	1	16	20.323	1.082	7.236	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	52.56
218	TTTTTTAGTT	9	10	4	4	16	20.323	0.639	3.203	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	52.53
219	TTAATTTAA	10	13	6	6	15	18.871	0.687	3.639	8	280	2	90	1	30	0	0	0	0	0	2	-10	30	52.51
220	CACTGCG	9	10	4	4	16	20.323	1.277	9.008	7	245	0	0	0	0	0	0	0	0	0	3	-15	23	52.33
221	AAAATTAATTTTT	5	6	0	0	16	20.323	0.69	3.669	0	0	4	180	0	0	0	0	0	0	0	2	-10	28.333	52.32
222	ATATGTGA	8	10	5	5	11	13.065	1.082	7.236	8	280	1	45	0	0	0	0	0	0	0	1	-5	32	52.30
223	TATTCCGT	7	7	2	2	15	18.871	1.213	8.426	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	52.30
224	CTTTTGAC	10	10	5	5	15	18.871	1.213	8.426	5	175	2	90	0	0	0	0	0	0	0	3	-15	25	52.30
225	ATTGTTTTTC	8	8	1	1	21	27.581	0.94	5.946	3	105	1	45	0	0	1	15	0	0	0	3	-15	18.75	52.28
226	TTGGGAGA	9	10	4	4	16	20.323	1.04	6.849	5	175	2	90	0	0	0	0	0	0	0	3	-15	25	52.17
227	GATTTATG	9	10	4	4	16	20.323	1.04	6.849	5	175	2	90	0	0	0	0	0	0	0	3	-15	25	52.17
228	AGTTTAAAT	9	11	4	4	17	21.774	0.965	6.169	4	140	3	135	0	0	0	0	1	5	0	3	-15	24.091	52.03
229	AAATGTAAT	8	10	5	5	11	13.065	0.937	5.913	7	245	2	90	0	0	0	0	0	0	0	1	-5	33	51.98
230	ATCACACA	7	9	2	2	17	21.774	0.974	6.254	4	140	2	90	0	0	0	0	0	0	0	3	-15	23.889	51.92
231	TGCAGCTG	7	7	2	2	15	18.871	1.321	9.408	5	175	0	0	0	0	0	0	0	0	0	2	-10	23.571	51.85
232	AATCATTAT	7	9	2	2	17	21.774	0.965	6.169	4	140	2	90	0	0	0	0	0	0	0	3	-15	23.889	51.83
233	TCGGCGAA	6	8	2	2	14	17.419	1.321	9.408	6	210	0	0	0	0	0	0	0	0	0	2	-10	25	51.83
234	AATGCATC	9	14	6	6	14	17.419	1.321	9.408	10	350	0	0	0	0	1	15	0	0	0	3	-15	25	51.83
235	AATCGGCG	4	6	0	0	14	17.419	1.321	9.408	2	70	2	90	0	0	0	0	0	0	0	2	-10	25	51.83
236	AAATCGGCG	4	6	0	0	14	17.419	1.311	9.316	2	70	2	90	0	0	0	0	0	0	0	2	-10	25	51.73
237	GGCCGTC	8	8	3	3	15	18.871	1.004	6.526	5	175	1	45	0	0	0	0	0	0	0	2	-10	26.25	51.65
238	GTATTTAAATA	4	6	0	0	14	17.419	0.935	5.894	5	175	0	0	0	0	0	0	0	0	0	1	-5	28.333	51.65
239	TACCGTC	8	9	3	3	16	20.323	1.277	9.008	4	140	1	45	1	30	0	0	0	0	0	3	-15	22.222	51.55
240	ATTTCTTTT	8	10	3	4	15	18.871	0.684	3.61	6	210	2	90	0	0	0	0	0	0	0	2	-10	29	51.48
241	ATTTTTATC	9	10	4	4	16	20.323	0.849	5.111	4	140	3	135	0	0	0	0	0	0	0	3	-15	26	51.43
242	TTTTATAAAA	6	8	2	2	14	17.419	0.693	3.696	7	245	0	0	0	0	0	0	0	0	0	1	-5	30	51.11

		Positive			Negative			Complexity		> 1kb	>1kb val	< 1kb	1kb value	Introns						Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	>3rd intron		exon_value	Pos_Score	Total Score
243	ATCAATAAA	9	9	3	3	18	23.226	0.849	5.111	5	175	1	45	0	0	0	0	0	0	0	3	-15	22.778	51.11
244	AAAAAATTTTTTT	2	7	0	0	11	13.065	0.69	3.669	1	35	4	180	1	30	0	0	0	0	0	1	-5	34.286	51.02
245	AAGAATCAA	8	8	2	2	18	23.226	1.003	6.512	4	140	1	45	0	0	0	0	0	0	0	3	-15	21.25	50.99
246	AACGTTGA	7	8	2	2	16	20.323	1.321	9.408	4	140	1	45	0	0	0	0	0	0	0	3	-15	21.25	50.98
247	TGGAAATAT	6	7	1	1	16	20.323	1.061	7.041	2	70	2	90	0	0	1	15	0	0	0	2	-10	23.571	50.94
248	GTTAAAATT	12	13	6	6	19	24.677	0.965	6.169	6	210	1	45	1	30	0	0	0	0	0	5	-25	20	50.85
249	TGAAAATTTC	7	9	4	4	11	13.065	1.194	8.249	6	210	0	0	2	60	0	0	0	0	0	1	-5	29.444	50.76
250	GCAACGAAA	5	6	1	1	13	15.968	0.995	6.442	5	175	0	0	0	0	0	0	0	0	0	1	-5	28.333	50.74
251	ATTGGGAGA	5	6	0	0	16	20.323	1.061	7.041	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	50.70
252	GAATACCA	7	8	4	4	10	11.613	1.213	8.426	5	175	1	45	1	30	0	0	0	0	0	1	-5	30.625	50.66
253	CATTGTAAA	6	6	1	1	15	18.871	1.215	8.443	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	50.65
254	GCTAGC	10	12	4	4	20	26.129	1.33	9.487	6	210	0	0	0	0	0	0	0	0	0	6	-30	15	50.62
255	GAGAGGC	6	6	2	2	12	14.516	0.956	6.085	4	140	1	45	0	0	0	0	0	0	0	1	-5	30	50.60
256	TTTCGTGA	9	13	6	6	13	15.968	1.213	8.426	4	140	0	0	7	210	0	0	0	0	0	2	-10	26.154	50.55
257	CTAATTTG	8	9	4	4	13	15.968	1.213	8.426	7	245	0	0	0	0	0	0	0	0	0	2	-10	26.111	50.50
258	GATATCAT	7	8	2	2	16	20.323	1.255	8.812	5	175	0	0	0	0	0	0	1	5	0	2	-10	21.25	50.38
259	CTTCATGA	7	7	3	3	12	14.516	1.321	9.408	3	105	2	90	0	0	0	0	0	0	0	2	-10	26.429	50.35
260	TTCACTGAA	4	7	0	0	15	18.871	1.311	9.316	3	105	1	45	0	0	1	15	0	0	0	2	-10	22.143	50.33
261	ATGAATTCA	7	8	4	4	10	11.613	1.215	8.443	7	245	0	0	0	0	0	0	0	0	0	1	-5	30	50.06
262	ACATTGAC	6	8	3	3	11	13.065	1.321	9.408	4	140	2	90	0	0	0	0	0	0	0	2	-10	27.5	49.97
263	TAATTTTTTT	11	14	6	6	18	23.226	0.474	1.703	5	175	4	180	0	0	1	15	0	0	0	4	-20	25	49.93
264	AAATTTTTATT	6	6	0	0	18	23.226	0.655	3.353	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	49.91
265	TGATGACA	9	13	7	7	10	11.613	1.321	9.408	6	210	4	180	0	0	0	0	0	0	0	3	-15	28.846	49.87
266	AAGTGCTC	7	9	3	3	14	17.419	1.386	10.003	4	140	1	45	1	30	0	0	0	0	0	3	-15	22.222	49.64
267	ATCCTAAT	6	7	2	2	13	15.968	1.082	7.236	3	105	2	90	0	0	0	0	0	0	0	2	-10	26.429	49.63
268	CGGTCTC	9	11	5	5	14	17.419	1.079	7.206	7	245	1	45	0	0	0	0	0	0	0	3	-15	25	49.63
269	AAAAAAATTAAA	9	9	2	2	21	27.581	0.451	1.488	4	140	1	45	0	0	1	15	0	0	0	3	-15	20.556	49.62
270	AATTTTGAAAG	7	7	1	1	18	23.226	1.036	6.817	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	49.33
271	ACACCGGA	5	6	0	0	16	20.323	1.082	7.236	4	140	0	0	0	0	0	0	0	0	0	2	-10	21.667	49.22
272	GTACAAAAA	8	10	4	5	12	14.516	1.003	6.512	7	245	1	45	0	0	0	0	0	0	0	2	-10	28	49.03
273	TGTTTCTA	8	10	6	6	8	8.71	1.074	7.157	7	245	2	90	0	0	0	0	0	0	0	1	-5	33	48.87
274	AATTGAGG	6	7	3	3	10	11.613	1.082	7.236	4	140	1	45	1	30	0	0	0	0	0	1	-5	30	48.85
275	TGAGATCA	7	9	3	3	14	17.419	1.321	9.408	6	210	0	0	0	0	0	0	0	0	0	3	-15	21.667	48.49
276	GATGAATG	6	9	3	4	10	11.613	1.082	7.236	4	140	3	135	0	0	0	0	0	0	0	2	-10	29.444	48.29
277	CGGTCTCG	7	8	3	3	13	15.968	1.082	7.236	6	210	0	0	0	0	0	0	0	0	0	2	-10	25	48.20
278	TTCTGTTTT	10	10	3	3	21	27.581	0.684	3.61	3	105	2	90	0	0	0	0	0	0	0	5	-25	17	48.19
279	TTTTTTTTGGA	6	9	2	2	15	18.871	0.76	4.3	3	105	3	135	0	0	0	0	0	0	0	3	-15	25	48.17
280	TGAACATG	8	9	4	4	13	15.968	1.321	9.408	5	175	1	45	0	0	0	0	0	0	0	3	-15	22.778	48.15
281	TTTCAACAAAAA	6	6	3	3	9	10.161	0.96	6.12	3	105	2	90	0	0	0	0	0	0	0	1	-5	31.667	47.95
282	AAAATCGAAAA	9	14	6	7	12	14.516	0.886	5.447	9	315	2	90	0	0	0	0	0	0	0	3	-15	27.857	47.82
283	TGATATCTT	6	6	0	0	18	23.226	1.149	7.844	2	70	1	45	0	0	0	0	0	0	0	3	-15	16.667	47.74
284	TGTAAAGA	10	11	5	5	16	20.323	1.04	6.849	2	70	4	180	0	0	0	0	0	0	0	5	-25	20.455	47.63
285	TGTGAGCA	7	7	2	2	15	18.871	1.321	9.408	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	47.56
286	CACTTTTTC	8	10	5	5	11	13.065	0.937	5.913	7	245	0	0	1	30	1	15	0	0	0	1	-5	28.5	47.48
287	CAACGTTT	10	11	5	6	14	17.419	1.321	9.408	7	245	0	0	0	0	0	0	0	0	0	4	-20	20.455	47.28
288	TCATAATAA	8	8	3	3	15	18.871	0.937	5.913	5	175	0	0	0	0	1	15	0	0	0	2	-10	22.5	47.28
289	ATATATATAT	3	4	1	1	7	7.258	0.693	3.696	2	70	1	45	1	30	0	0	0	0	0	0	0	36.25	47.20
290	TAAAGTTTG	6	6	3	3	9	10.161	1.061	7.041	4	140	1	45	0	0	0	0	0	0	0	1	-5	30	47.20
291	AGTTTATCT	5	6	1	1	13	15.968	1.149	7.844	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	47.15

		Positive			Negative			Complexity		> 1kb	>1kb val	< 1kb	1kb value	Introns						>3rd	Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	intron		exon_value	Pos_Score	Total Score
292	ATTTCCAAT	10	11	6	6	13	15.968	1.061	7.041	8	280	0	0	0	0	0	0	0	0	0	3	-15	24.091	47.10
293	GCAACCG	5	9	2	2	13	15.968	1.079	7.206	4	140	2	90	0	0	0	0	0	0	0	3	-15	23.889	47.06
294	ATTGTAAC	5	7	1	1	14	17.419	1.255	8.812	4	140	0	0	0	0	1	15	0	0	0	2	-10	20.714	46.95
295	TAACAAATT	6	7	2	2	13	15.968	0.937	5.913	4	140	1	45	0	0	0	0	0	0	0	2	-10	25	46.88
296	GATCGATT	7	7	2	2	15	18.871	1.321	9.408	1	35	2	90	0	0	0	0	0	0	1	3	-15	18.333	46.61
297	CTGTTGAA	10	11	6	7	11	13.065	1.321	9.408	5	175	2	90	0	0	1	15	0	0	0	3	-15	24.091	46.56
298	CCTATCT	7	10	5	5	9	10.161	1.004	6.526	7	245	2	90	1	30	0	0	0	0	0	2	-10	29.583	46.27
299	TTTTGGGAA	4	9	1	1	14	17.419	1.061	7.041	1	35	4	180	0	0	0	0	0	0	0	4	-20	21.667	46.13
300	TTTCCAGCA	7	7	1	1	18	23.226	1.311	9.316	2	70	1	45	0	0	0	0	0	0	0	4	-20	13.571	46.11
301	GTGAGAGC	5	6	1	1	13	15.968	1.213	8.426	4	140	0	0	0	0	0	0	0	0	0	2	-10	21.667	46.06
302	ACCGGAG	8	9	3	3	16	20.323	1.079	7.206	4	140	1	45	0	0	0	0	0	0	0	4	-20	18.333	45.86
303	TTTTTCTTTC	6	7	2	2	13	15.968	0.5	1.942	2	70	3	135	0	0	0	0	0	0	0	2	-10	27.857	45.77
304	TCACAAAAAA	7	8	3	3	13	15.968	0.802	4.684	6	210	0	0	0	0	0	0	0	0	0	2	-10	25	45.65
305	TTTTGGGAAA	2	6	0	0	10	11.613	1.089	7.297	1	35	3	135	0	0	0	0	0	0	0	2	-10	26.667	45.58
306	TTCATGATT	5	6	1	1	13	15.968	1.149	7.844	1	35	1	45	2	60	0	0	0	0	0	2	-10	21.667	45.48
307	AAGGGCT	6	9	2	2	15	18.871	1.277	9.008	5	175	0	0	0	0	0	0	0	0	0	4	-20	17.222	45.10
308	TTAATTTATT	9	9	3	4	16	20.323	0.611	2.947	3	105	2	90	0	0	1	15	0	0	0	3	-15	21.667	44.94
309	CTGCACC	7	8	2	2	16	20.323	1.154	7.887	3	105	1	45	0	0	0	0	0	0	0	4	-20	16.25	44.46
310	AAAATTTCAAAA	5	10	2	4	10	11.613	0.824	4.886	6	210	1	45	1	30	0	0	0	0	0	2	-10	27.5	44.00
311	ATATCTTA	8	9	4	4	13	15.968	0.974	6.254	6	210	0	0	0	0	0	0	0	0	0	3	-15	21.667	43.89
312	TGAAAAATGA	6	6	2	2	12	14.516	0.95	6.035	3	105	1	45	0	0	0	0	0	0	0	2	-10	23.333	43.88
313	TTAGAAATA	5	6	1	1	13	15.968	0.937	5.913	4	140	0	0	0	0	0	0	0	0	0	2	-10	21.667	43.55
314	CAATTTTTTT	8	10	6	6	8	8.71	0.802	4.684	4	140	3	135	1	30	0	0	0	0	0	2	-10	29.5	42.89
315	TTTTTTCGTT	6	7	1	1	16	20.323	0.639	3.203	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	42.81
316	CATCGACA	5	7	0	0	17	21.774	1.255	8.812	3	105	0	0	0	0	0	0	0	0	0	4	-20	12.143	42.73
317	AATTTAATTTT	8	8	4	4	12	14.516	0.655	3.353	2	70	2	90	1	30	1	15	0	0	0	2	-10	24.375	42.24
318	AATGAATAT	7	9	3	3	14	17.419	0.937	5.913	2	70	2	90	1	30	0	0	0	0	0	4	-20	18.889	42.22
319	GAATTACA	6	8	2	2	14	17.419	1.213	8.426	3	105	1	45	0	0	0	0	0	0	0	4	-20	16.25	42.10
320	TTCCGTTTTT	7	7	1	1	18	23.226	0.802	4.684	2	70	1	45	0	0	0	0	0	0	0	4	-20	13.571	41.48
321	TGATCATG	9	9	5	5	12	14.516	1.321	9.408	5	175	0	0	0	0	0	0	0	0	0	4	-20	17.222	41.15
322	CACTGAAAC	5	6	1	1	13	15.968	1.215	8.443	2	70	1	45	0	0	0	0	0	0	0	3	-15	16.667	41.08
323	CTGTTTTG	10	11	6	7	11	13.065	0.9	5.58	5	175	2	90	0	0	0	0	0	0	0	4	-20	22.273	40.92
324	GCATCAC	10	12	8	8	8	8.71	1.277	9.008	6	210	1	45	1	30	0	0	0	0	0	4	-20	22.083	39.80
325	CGTCCGGT	3	8	0	0	14	17.419	1.082	7.236	4	140	0	0	0	0	0	0	0	0	0	4	-20	15	39.65
326	AATTGTTTC	7	7	3	3	12	14.516	1.149	7.844	3	105	0	0	1	30	0	0	0	0	0	3	-15	17.143	39.50
327	CAGTTTCT	7	10	6	6	6	5.806	1.213	8.426	5	175	2	90	0	0	0	0	0	0	0	3	-15	25	39.23
328	AAATTTTTAT	7	10	4	5	10	11.613	0.673	3.512	6	210	1	45	0	0	0	0	0	0	0	3	-15	24	39.13
329	ACATAATC	7	8	2	2	16	20.323	1.04	6.849	0	0	2	90	1	30	0	0	0	0	0	5	-25	11.875	39.05
330	TTTGGATTT	7	9	2	3	15	18.871	0.849	5.111	2	70	2	90	0	0	0	0	0	0	0	5	-25	15	38.98
331	ATATAATTTA	6	7	2	2	13	15.968	0.693	3.696	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	38.95
332	TTAATTTAT	11	12	6	7	14	17.419	0.637	3.18	4	140	2	90	0	0	1	15	0	0	0	5	-25	18.333	38.93
333	GATCTCAA	6	8	2	3	12	14.516	1.321	9.408	4	140	0	0	0	0	0	0	0	0	0	4	-20	15	38.92
334	TATAATTTA	7	12	4	4	14	17.419	0.687	3.639	3	105	3	135	0	0	0	0	0	0	0	6	-30	17.5	38.56
335	GAAGTTTGA	5	7	2	2	11	13.065	1.099	7.385	4	140	0	0	0	0	0	0	0	0	0	3	-15	17.857	38.31
336	CTGTCATT	8	9	4	4	13	15.968	1.213	8.426	3	105	1	45	0	0	0	0	0	0	0	5	-25	13.889	38.28
337	TCCGGTT	7	10	3	3	15	18.871	1.079	7.206	3	105	1	45	0	0	0	0	0	0	0	6	-30	12	38.08
338	AAATTTAATTT	9	11	5	5	14	17.419	0.689	3.658	0	0	3	135	2	60	1	15	0	0	0	5	-25	16.818	37.90
339	AATGGTGA	8	8	3	3	15	18.871	1.082	7.236	2	70	1	45	0	0	0	0	0	0	0	5	-25	11.25	37.36
340	TTGAAGTC	7	8	4	4	10	11.613	1.321	9.408	3	105	1	45	0	0	0	0	0	0	0	4	-20	16.25	37.27

		Positive			Negative			Complexity		> 1kb	>1kb_val	< 1kb	1kb_value	Introns							Exon			
#	Motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Freq_Val	Freq_Score	Value	Score					1st intron	Intron_value	2nt intron	Intron_val	3rd intron	Intron_va	>3rd intron		exon_value	Pos_Score	Total Score
341	CAATTTGAA	7	9	2	4	13	15.968	1.215	8.443	4	140	0	0	0	0	0	0	0	0	0	5	-25	12.778	37.19
342	TTGATTTGT	8	9	4	4	13	15.968	0.849	5.111	2	70	1	45	1	30	1	15	0	0	0	4	-20	15.556	36.63
343	GGCCAAAA	8	10	4	4	14	17.419	1.04	6.849	3	105	1	45	0	0	0	0	0	0	0	6	-30	12	36.27
344	TTCTGATTT	8	10	4	4	14	17.419	1.003	6.512	3	105	1	45	0	0	0	0	0	0	0	6	-30	12	35.93
345	TTTTTTTTTTGA	6	7	2	3	11	13.065	0.536	2.265	3	105	1	45	0	0	0	0	0	0	0	3	-15	19.286	34.62
346	TATCTTAA	8	10	5	5	11	13.065	0.974	6.254	5	175	0	0	0	0	0	0	0	0	0	5	-25	15	34.32
347	AAAAAATATAT	4	7	2	2	9	10.161	0.586	2.72	0	0	3	135	1	30	0	0	0	0	0	3	-15	21.429	34.31
348	CATCACCA	4	8	3	3	7	7.258	0.974	6.254	3	105	1	45	1	30	0	0	0	0	0	3	-15	20.625	34.14
349	TTTTTTTTTTG	7	11	4	5	11	13.065	0.287	0	4	140	2	90	0	0	0	0	0	0	0	4	-20	21	34.06
350	TAGAAATAT	7	8	4	4	10	11.613	0.937	5.913	3	105	1	45	0	0	0	0	0	0	0	4	-20	16.25	33.78
351	ATTTTTTTTTT	6	9	3	3	12	14.516	0.287	0	2	70	2	90	0	0	0	0	0	0	0	5	-25	15	29.52

Motifs identified from the muscle data set.

			Positive set		Negative set			Position			Complexity		Total
#	Motif	Mus motif	Frag Freq	Tot Freq	Frag Freq	Tot Freq	Score	>1kb	<1kb	Score	Value	Score	Score
1	**tcccgggа**	Y	3	3	0	0	65.7	0	3	15.0	1.255	12.9	93.62
2	cgggatcc	Y	5	5	0	0	68.6	3	2	12.0	1.255	12.9	93.48
3	attagtgca		5	5	0	0	68.6	4	1	11.0	1.311	13.8	93.36
4	atgatcatca		4	4	0	0	67.1	2	2	12.5	1.280	13.3	92.94
5	gatcatcaa		4	4	0	0	67.1	2	2	12.5	1.273	13.2	92.83
6	**caattgctg**		5	6	0	0	66.2	4	2	11.7	1.369	14.7	92.58
7	**agcaattgc**		3	4	0	0	66.2	3	1	11.7	1.369	14.7	92.58
8	**cagcaattgc**		3	4	0	0	66.2	3	1	11.7	1.366	14.7	92.54
9	cctaggag		4	4	0	0	67.1	3	1	11.3	1.321	14.0	92.35
10	**atatgcta**		6	6	1	1	68.6	5	1	10.8	1.255	12.9	92.31
11	gcaattgc		8	10	6	6	65.2	6	4	12.0	1.386	15.0	92.24
12	atcagtaat		4	4	0	0	67.1	2	2	12.5	1.215	12.3	91.90
13	**attagtgc**		5	6	0	0	66.2	4	2	11.7	1.321	14.0	91.81
14	**agcaattg**		4	5	2	2	64.8	2	3	12.8	1.321	14.0	91.50
15	gatgatcatc		2	2	0	0	64.3	1	1	12.5	1.342	14.3	91.08
16	tgatgatcatca		2	2	0	0	64.3	1	1	12.5	1.330	14.1	90.88
17	**cgggatc**	Y	8	8	5	5	65.7	5	3	11.9	1.277	13.3	90.84
18	aactttgcaa		4	4	0	0	67.1	4	0	10.0	1.280	13.3	90.44
19	**gctctgcc**		6	6	2	2	67.1	2	4	13.3	1.040	9.4	89.92
20	gcgggagc	Y	5	5	0	0	68.6	1	4	14.0	0.900	7.2	89.78
21	ggaaagcc		4	4	0	0	67.1	2	2	12.5	1.082	10.1	89.77
22	gatgatcatca		1	1	0	0	62.9	2	2	12.5	1.342	14.3	89.65
23	cccgggа	Y	7	7	4	4	65.7	0	7	15.0	1.004	8.9	89.59
24	atgatcatc		1	1	0	0	62.9	2	2	12.5	1.330	14.1	89.45
25	ctataca		5	6	1	1	67.6	4	2	11.7	1.079	10.1	89.36
26	agcaattgctg		2	2	0	0	64.3	4	0	10.0	1.373	14.8	89.08
27	agcaattgct		2	2	0	0	64.3	4	0	10.0	1.366	14.7	88.97
28	cagcaattgctg		2	2	0	0	64.3	2	0	10.0	1.330	14.1	88.38
29	gggagccc	Y	4	4	0	0	67.1	2	2	12.5	0.974	8.4	88.04
30	caatatcaa		5	5	1	1	67.1	3	2	12.0	0.995	8.7	87.87
31	**cccgcgg**	Y	9	9	3	3	70.0	2	7	13.9	0.683	3.7	87.61
32	caaacacga		4	4	0	0	67.1	2	2	12.5	0.937	7.8	87.44
33	cgggagc	Y	9	10	7	7	64.8	1	9	14.5	0.956	8.1	87.36
34	ccttccca		5	5	1	1	67.1	2	3	13.0	0.900	7.2	87.35
35	cgcgggag	Y	5	5	1	1	67.1	2	3	13.0	0.900	7.2	87.35
36	attcttcacta		4	4	0	0	67.1	4	0	10.0	1.067	9.9	87.03
37	cgcagac		10	10	8	9	63.8	4	6	13.0	1.079	10.1	86.88
38	atatatgat		4	4	0	0	67.1	3	1	11.3	0.965	8.2	86.64
39	tagtgaagaa		4	4	0	0	67.1	4	0	10.0	1.030	9.3	86.43

			Positive set		Negative set			Position			Complexity		Total
#	Motif	Mus_motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Score	>1kb	<1kb	Sscore	Value	Score	Score
40	attttaatttc		4	4	0	0	67.1	2	2	12.5	0.860	6.6	86.20
41	**cgctccc**	Y	4	5	0	0	67.6	2	3	13.0	0.796	5.5	86.16
42	aagaagaagc	#3	7	7	5	5	64.3	1	6	14.3	0.898	7.2	85.74
43	cgcggga	Y/#5	9	9	7	7	64.3	3	6	13.3	0.956	8.1	85.72
44	Ggataga		7	8	5	5	64.8	6	2	11.3	1.004	8.9	84.89
45	cttttctctc		4	4	0	0	67.1	1	3	13.8	0.673	3.6	84.45
46	tcccaca		8	8	6	6	64.3	5	3	11.9	0.956	8.1	84.26
47	agggcgg	#4	8	8	5	6	65.2	3	5	13.1	0.796	5.5	83.90
48	gctccc	Y	17	23	17	19	63.3	7	16	13.5	0.868	6.7	83.49
49	acacaaac		7	7	4	4	65.7	1	6	14.3	0.662	3.4	83.38
50	ctcccac		8	8	5	5	65.7	5	3	11.9	0.796	5.5	83.13
51	cacacaaa		12	13	9	9	66.2	4	9	13.5	0.662	3.4	83.03
52	**cccgcggg**	Y	4	4	2	2	64.3	1	3	13.8	0.693	3.9	81.92
53	acccccctc		4	4	0	0	67.1	4	0	10.0	0.736	4.6	81.71
54	gggggggca		4	4	0	0	67.1	4	0	10.0	0.736	4.6	81.71
55	ccgagg		10	11	12	12	59.0	3	8	13.6	1.011	9.0	81.67
56	cacccccctc		4	4	0	0	67.1	4	0	10.0	0.684	3.7	80.88
57	ggcgggga	Y/#5	10	11	10	10	61.9	4	7	13.2	0.796	5.5	80.63
58	acccaat		11	12	13	13	59.0	9	3	11.3	1.004	8.9	79.17
59	gaaatata		11	12	12	12	60.5	9	3	11.3	0.900	7.2	78.93
60	ccgcgg	Y	6	7	6	7	61.4	2	5	13.6	0.693	3.9	78.88
61	gaaaaaagaa		9	9	6	7	65.2	5	4	12.2	0.500	0.8	78.25
62	cccgcg	Y	13	16	14	14	61.4	5	11	13.4	0.637	3.0	77.84
63	acacgc		14	14	19	22	52.9	7	7	12.5	1.011	9.0	74.35
64	gcccgc	Y	10	12	13	13	58.1	6	6	12.5	0.637	3.0	73.57
65	aaatatac		11	12	15	15	56.2	12	0	10.0	0.900	7.2	73.40
66	atatctc		13	16	22	25	48.6	6	10	13.1	1.079	10.1	71.77
67	aacagag		12	15	21	21	50.0	4	11	13.7	0.956	8.1	71.76
68	gcacta		15	18	27	27	45.7	13	5	11.4	1.330	14.1	71.20
69	gcacaca		14	16	23	24	49.0	8	8	12.5	1.004	8.9	70.42
70	ctatcc		13	16	22	24	49.0	11	5	11.6	1.011	9.0	69.60
71	gatccc		15	17	24	33	45.2	12	5	11.5	1.242	12.7	69.41
72	aaggct		15	17	27	33	42.4	8	9	12.6	1.330	14.1	69.12
73	gccccc		10	11	13	14	57.1	7	4	11.8	0.451	0.0	68.95
74	ccgcgc	Y	11	13	18	19	51.9	4	9	13.5	0.637	3.0	68.34
75	atcccg		14	18	28	30	42.4	9	9	12.5	1.242	12.7	67.58
76	gaaaaaaga		11	11	17	19	51.9	6	5	12.3	0.530	1.3	65.44
77	atatatg		13	15	26	27	43.3	7	8	12.7	1.004	8.9	64.88
78	cagagc		16	21	31	34	41.0	9	12	12.9	1.099	10.4	64.20
79	acgcct		16	17	28	41	38.6	8	9	12.6	1.242	12.7	63.92

			Positive set		Negative set			Position			Complexity		Total
#	Motif	Mus_motif	Frag_Freq	Tot_Freq	Frag_Freq	Tot_Freq	Score	>1kb	<1kb	Sscore	Value	Score	Score
80	cgggag	Y	16	25	32	34	41.9	6	19	13.8	0.868	6.7	62.39
81	cacccc		14	15	22	22	50.5	12	3	11.0	0.451	0.0	61.47
82	aacatgt		15	16	33	37	34.3	10	6	11.9	1.277	13.3	59.41
83	aatatac		16	19	31	34	40.0	14	5	11.3	0.956	8.1	59.41
84	gtggga		16	22	31	36	40.5	13	9	12.0	0.868	6.7	59.20
85	ccttaa		17	19	33	42	35.2	8	11	12.9	1.099	10.4	58.52
86	cccctc		13	13	22	25	47.1	10	3	11.2	0.451	0.0	58.29
87	aaagcc		17	20	31	45	36.2	8	12	13.0	1.011	9.0	58.18
88	cccaca		17	20	31	32	42.4	10	10	12.5	0.637	3.0	57.86
89	cccgcc	Y/#5	15	19	30	31	41.4	4	15	13.9	0.451	0.0	55.37
90	gcagac		17	23	37	45	31.9	10	13	12.8	1.099	10.4	55.12
91	cacacaa		14	16	31	34	36.7	5	11	13.4	0.683	3.7	53.82
92	acacaaa		18	23	39	49	29.0	6	17	13.7	0.598	2.4	45.11
93	tggagg		19	25	42	48	28.6	8	17	13.4	0.637	3.0	44.95
94	cgagag	#2	19	31	48	57	21.4	9	22	13.5	1.011	9.0	43.97
95	agtgta		18	23	46	55	19.5	11	12	12.6	1.099	10.4	42.52
96	cctctc	#2	19	27	47	55	21.4	8	19	13.5	0.637	3.0	37.92
97	atagag		19	25	51	66	11.4	7	18	13.6	1.011	9.0	34.02
98	gtgtta		18	20	51	61	10.5	9	11	12.8	1.011	9.0	32.22
99	atgtgg		18	22	57	66	3.3	12	10	12.3	1.011	9.0	24.60
100	aaggga		18	21	59	68	0.0	8	13	13.3	0.693	3.9	17.13

Key

Y represents the presence of all motifs that are variations of the muscle *cis*-acting element (cccgCGGGagcccg) and in the same column the numbers represent the other motifs detected by Guhathakurta *et al.*, (2002b); #2 CTCTcaaaccc, #3 aAGAAGAagc, #4 TGGGcGGa and #5 ggGCGGGa. Emboldened motifs are those detected using the frequency cut-off defined in the genomic insert approach.