DE NOVO DETECTION OF REGULATORY ELEMENTS IN THE NEMATODE *CAENORHABDITIS ELEGANS*

by

MONICA CELIA SLEUMER

Bachelor of Science, University of British Columbia, 1999 (Biochemistry)
Bachelor of Science, University of British Columbia, 2001 (Computer Science)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Genetics)

The University of British Columbia

(Vancouver)

July 2009

Abstract

The availability of high-throughput gene expression data and completely sequenced genomes from eight species of nematodes has provided an opportunity to identify novel *cis*-regulatory elements in the promoter regions of *Caenorhabditis elegans* transcripts.

Motif discovery was performed in the promoter regions of genes expressed in the *C. elegans* intestine. We scanned the upstream regions of genes expressed in the intestine and ASE neurons for sequences similar to the binding sites of the transcription factors ELT-2 and CHE-1 respectively and showed that they are more likely to contain high-scoring matches to these binding sites than upstream regions of other genes.

To create the cisRED *C. elegans* database, we determined orthologues for *C. elegans* transcripts in *C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica*, *Pristionchus pacificus*, *Brugia malayi* and *Trichinella spiralis* using the WABA alignment algorithm. We pooled the upstream region of each transcript in *C. elegans* with the upstream regions of its orthologues and identified conserved DNA sequence elements by *de novo* motif discovery. In total, we discovered 158,017 novel conserved motifs upstream of 3847 *C. elegans* transcripts for which three or more orthologues were available, and identified 82% of 44 experimentally validated regulatory elements from the ORegAnno TFBS database. We annotated 26% of the motifs as similar to known binding sequences of transcription factors from the ORegAnno, TRANSFAC and JASPAR databases. This is the first catalogue of annotated conserved upstream elements for nematodes and can be used to find putative regulatory elements, improve gene models, discover novel RNA genes, and understand the evolution of transcription factors and their binding sites in phylum Nematoda.

We placed the cisRED motifs into groups based on sequence similarity and identified a series of motif groups that are associated with genes that have significant functional associations. Fifteen of the groups are specifically associated with ribosomal protein genes. Eight of these are extensions of the canonical *C. elegans* trans-splice acceptor site; two are similar to binding sites of transcription factors in other species. One was tested for regulatory function in a series of GFP expression experiments and was shown to be involved in pharyngeal expression.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ATG	Translation start site
bp	base pairs
CDF	Cumulative distribution function
ChIP	Chromatin immunoprecipitation
DBD	DNA binding domain
EMSA	Electrophoretic mobility shift assay
FACS	Fluorescence-activated cell sorter
GFP	Green fluorescent protein
GO	Gene ontology
IC	Information content
IUPAC	International union of pure and applied chemistry
KOG	Eukaryotic clusters of orthologous groups
NHR	Nuclear hormone receptor
PCR	Polymerase chain reaction
PFM	Position frequency matrix
SAGE	Serial analysis of gene expression
SELEX	Systematic evolution of ligands by exponential enrichment
TESS	Transcriptional element search system
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site
UTR	Untranslated region

Acknowledgements

I would like to thank my graduate supervisor, Steven Jones, for the opportunity to pursue graduate studies at the Genome Sciences Centre and for supporting all of my research and travel endeavours. Thanks also to my thesis committee members: Robert Holt, Don Moerman, and Mark Wilkinson, for your guidance, and to the Director of the Genetics Graduate Program, Hugh Brock, for operating a well-organized department. I am grateful to the Michael Smith Foundation for Health Research for providing three years of stipend funding, plus funding for travel and equipment.

Thank you to my current and former co-workers at the GSC, particularly Misha Bilenky, Gordon Robertson, Nina Thiessen, An He, Tamara Astakhova, Maik Hassel, Greg Taylor, Diana Palmquist Harvey, Kim Wong, Melis Dagpinar, Keven Lin, and co-op students Wenjia Pan and Maggie Xin Zhang, and to my fellow graduate students, especially Obi Griffith, Carri-lyn Mead, Sorana Morrissy, Heesun Shin, Malachi Griffith, Erin Pleasance, Stephen Montgomery, Yvonne Li, Noushin Farnoud, and Claire Hou.

I would like to thank my collaborators: Jim McGhee, Oliver Hobert, David Baillie, and Allan Mah, for inviting me to contribute to your publications and for contributing to mine. Thank you to Andrew Smith for providing a custom-built version of DME. I would also like to thank all members of the Vancouver worm research community, particularly Ann Rose and Nigel O'Neil, for useful talks, meetings, and discussions.

Finally, I would like to thank Greg Vatcher, my parents Bernhard and Foyita Sleumer, and my sisters Karen Little and Nora Sleumer for your loving support and encouragement over the years.

To my parents

Co-authorship Statement

Chapter 2 draws from two separate publications to which my contributions were similar. For both studies, SAGE libraries were created by the British Columbia *C. elegans* Gene Expression Consortium consisting of the David Baillie and Donald Moerman laboratories and Canada's Michael Smith Genome Sciences Centre. The text of the chapter is entirely new and was written solely by me.

For the Intestinal study, SAGE data analysis was performed by Dr. James McGhee and members of his lab. The 74 specifically-expressed transcripts and list of experimentally validated ELT-2 binding sites were identified by Dr. James McGhee. Motif discovery, clustering, position frequency matrix generation, logo generation, motif scanning, and cumulative distribution function analysis were performed by myself in collaboration with Dr. Gordon Robertson and Dr. Mikhail Bilenky of the Genome Sciences Centre. The program that generates the sequence logos was written by Dr. Mikhail Bilenky. All sequence logos and graphs that appear in Chapter 2 were made by me, but some of the logos and graphs that appeared in the McGhee *et al.* publication were made by Dr. Mikhail Bilenky.

For the ASE neuron study, the image of the worm embryo expressing GFP in the ASE neuron (Figure 2.2, right) was taken by Dr. Oliver Hobert. CHE-1 was identified as the primary TF involved in regulation of genes in ASE neurons by Dr. Oliver Hobert. Scanning deletion mutagenesis experiments were performed by Dr. John F. Etchberger of the Hobert lab. SAGE data analysis was performed by Adam Lorch of the Moerman lab. The CHE-1 site position frequency matrix was made by Dr. Oliver Hobert. Logo generation, motif scanning, and cumulative distribution function analysis were performed by me.

Chapter 3 is a reprint of a paper published in Nucleic Acids Research with minor formatting and editorial changes. I was the primary author and was responsible for all analysis, text, figures, and tables. Dr. Mikhail Bilenky and Dr. Gordon Robertson designed and created the cisRED genome analysis pipeline for mammalian genomes, An He made significant improvements to the pipeline, and I adapted the pipeline for nematode genomes. Nina Thiessen improved and maintained the cisRED database and web interface. Dr. Steven Jones conceived the concept of the cisRED genome analysis pipeline and supervised the study.

Chapter 4 is a draft of a paper to be published. I was the primary author and was responsible for all analysis, text, figures, tables, and primer design. Dr. Allan Mah of the Baillie lab generated and injected the GFP constructs, maintained and photographed the worms, interpreted the GFP expression patterns, and performed the EMSAs.

1 Introduction

The availability of large-scale biological data has resulted in the need for bioinformatic analysis and has simultaneously provided an opportunity to gain an unprecedented insight into gene regulation. One hundred eukaryotic genomes have been completely sequenced, including eight nematode genomes [1]. Gene prediction and annotation programs allow us to interpret the genome sequences and relate findings in one genome directly to all of the others. High-throughput molecular biology techniques allow us to measure expression levels of all genes in any tissue. Finally, advanced computer algorithms and CPU clusters make it possible to sort through all of these data and find patterns that would take humans too long to find. The goal of this research is to increase our understanding of gene regulation in *C. elegans* through an extensive survey of upstream regions – searching for elements that are conserved among coexpressed genes, orthologous genes, and genes with related biological function.

1.1 Gene Regulation

All cells in an organism contain the same DNA sequence, but different cells use different genes at different times. Gene regulation is the process by which gene expression is turned on or off in a particular cell at a particular time. Gene regulation occurs at three levels: pre-transcriptional, by way of chromatin organization, histone modification, and DNA methylation; cotranscriptional, by way of the activation or repression of transcription and the processing and transport of the message to the ribosome; and post-transcriptional, by way of translation efficiency, mRNA stability, and binding of the message by proteins and other RNAs. The three levels are interdependent and may overlap somewhat; for instance, chromatin remodelling is known to occur during transcription, and cotranscriptional processing has an impact on mRNA stability. Each of these components of gene regulation, not to mention the interactions of all components simultaneously, are only partially understood and are under active research.

1.1.1 Pre-transcriptional Regulation

The broadest and most general level of gene regulation occurs via organization of chromatin, by which broad genomic regions are made available or unavailable to transcription. In the nucleus, DNA is wrapped around blocks of eight histone proteins called nucleosomes. Each nucleosome has 147 base pairs (bp) of DNA wrapped around it, and the nucleosomes are typically spaced about 50 bp apart [2]. Post-translational modifications of the histones' N-terminal tails, such as acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, and ribosylation have a profound impact on the packing structure of the nucleosomes and thereby influence the availability of the DNA for transcription [3]. Regions of the genome containing genes under active transcription in the cell are in the open euchromatin formation, characterized by methylation of specific lysine residues on histone H3 and other specific modifications. Regions not under active transcription, including silent genes, centromeres, telomeres, inactivated X chromosomes, and repeat sequences are in the tightly furled heterochromatin formation, characterized by acetylation and methylation of different lysine residues than that of euchromatin [2]. Some short intergenic regions are not bound by histones at all in order to allow DNA binding proteins such as transcription factors (TFs) to bind freely. These are the regions that are hypersensitive to DNase I (see section 1.2.1 below) [2].

The genomic pattern of chromatin organization changes through different stages of development and differs between tissue types [2]. Alteration of the pattern plays an important role throughout embryonic development and sexual maturation and is catalyzed by enzymes such as histone-acetyltransferases, deacetylases, and methyltransferases [3]. In mammals, direct methylation of the DNA is used to specify the chromatin pattern through multiple cell divisions in a process called genomic imprinting [2]. However, DNA methylation has not been observed in nematodes [4]; this may be due to the smaller number of cell types and the less complicated embryonic development of nematodes which requires less elaborate mechanisms of gene regulation.

1.1.2 Cotranscriptional Regulation

Transcription is a complex process in which genes are copied from DNA to RNA by a DNA-dependent RNA polymerase enzyme. Eukaryotic cells contain three types of RNA polymerase: RNA polymerase I transcribes mainly ribosomal RNA genes, RNA polymerase II transcribes protein-coding genes, and RNA polymerase III transcribes small nuclear and transfer RNA genes [5]. Because it performs the majority of the transcription in the cell and has undergone the most study, this discussion will focus on the regulation of transcription by RNA Polymerase II. Aspects that impact the efficiency of transcription include: the binding of specific TFs to enhancers near the gene, the binding of general TFs and cofactors to the promoter region of the gene, recruitment of the RNA polymerase enzyme itself, initiation of RNA synthesis, and transition to the faster elongation phase of RNA synthesis. Additionally, while the RNA is being synthesized, it undergoes cotranscriptional modifications such as 5' capping, cis-splicing, and polyadenylation [6]. Finally, after transcription is complete, the nascent RNA molecule is bound by stability-enhancing proteins and is transported out of the nucleus to undergo translation at the ribosome.

This entire process is regulated on many levels due to its complexity, however, the activation of the transcriptional machinery by gene-specific TFs is the most fundamental regulation mechanism [5]. The order in which the transcriptional components assemble is still under debate, but it is clear that the binding of the specific TFs to their DNA binding sites is the rate-limiting step. Here, we describe the transcriptional components in order of importance and specificity.

The binding of specific TFs to DNA sequences upstream of a gene is the most essential step of gene activation. Specific TFs may act on various levels to control gene expression, including the spatial, tissue, and individual cell levels [7]. A single primary TF may be involved in the expression of many genes in a particular physiological region of an organism (the "spatial" level). For example the winged helix factor PHA-4 is responsible for most pharyngeal gene expression, even though more than one cell type is involved [8]. A primary TF may also be involved in the expression of genes in a specific tissue type. For example, the GATA-type zinc finger TF ELT-2 is involved in the expression of most or all genes in *C. elegans* intestinal cells [9]. Finally, TFs may act on the expression of genes in only a single cell type. For example, the Paired-type homeobox factor CEH-10 and the LIM-type homeobox factor TTX-3 specify the expression of certain genes in *C. elegans* AIY interneurons [10].

There are several mechanisms by which TFs or other proteins repress or prevent gene expression instead of activating it. For example, a repressing protein might lack an activation domain and also have a stronger affinity for the same DNA sequence as the activating factor, thereby blocking gene activation whenever the repressor is present. Similarly, a repressing protein might dimerize with an activating TF and prevent its ability to activate gene expression via a protein domain that physically blocks the DNA binding domain or activation domain of the activating TF [11]. Some proteins act as both activators and repressors depending on where they are bound in the genome, whether they are phosphorylated, or whether they are bound to a specific ligand or cofactor [11]. One example of a repressor in *C. elegans* is the PIE-1 protein, which is an RNA binding protein that represses transcription by blocking the phosphorylation of the RNA Polymerase II carboxy-terminal domain, thereby preventing transition to the elongation phase of transcription [5].

The next important step in the initiation of transcription is formation of the pre-initiation complex. The pre-initiation complex is really a complex of subcomplexes, consisting of the RNA Polymerase II subcomplex, the TFIID subcomplex, and the mediator subcomplex, and is regulated on many levels due to the interchangeable parts in each complex [5]. Each of these three subcomplexes will be discussed in turn.

The RNA Polymerase II complex is a large enzyme that physically traverses the DNA strand and synthesizes the RNA strand. It has a large number of subunits that are called TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH, some of which are themselves smaller complexes [12] (Figure 1.1). The three-dimensional conformation of the entire complex is optimized to stabilize a short bubble of melted DNA, and also contains a groove for the nascent RNA strand [13]. The largest subunit of the complex has a carboxy-terminal domain that must be phosphorylated in order to make the transition from initiation to elongation. The carboxy-terminal domain also binds the components that are responsible for the cotranscriptional modifications discussed below. Other subunits of the RNA Polymerase II complex include chromatin remodellers, which modify the histones to a conformation that is suitable for transcription to proceed through the entire gene [5,14].

The TFIID complex is composed of the TATA binding protein, the cyclin-dependent kinase subunit, and 13 or 14 TATA binding protein-associated factors (Figure 1.1) [5]. The exact roles and structures of all of the subunits of the complex are still under investigation, but the overal structure of the complex has been shown to be conserved among yeast, *C. elegans*, *Drosophila*, and mammals [12]. The primary purpose of the TFIID complex may be to recognize the transcription start site (TSS) and anchor the RNA Polymerase II complex to the appropriate location on the DNA prior to transcriptional initiation. Several of the subunits of the complex are similar in structure to histones and may bind DNA in the same way [12]. Some eukaryotic promoters have a TATA sequence to which the TATA binding protein can bind; but many promoters lack a TATA sequence, and for others a TATA sequence has been predicted without experimental evidence to validate its function [7]. For those promoters that lack a TATA sequence, the TFIID complex may be recruited to the promoter by specific TFs, or else by modified histones in the promoter that are in the transcriptionally active state; three of the subunits of this complex are predicted to contain histone binding domains [12]. Once all of the necessary factors are in their appropriate locations on the gene promoter, including specific TFs and the RNA Polymerase II complex, and transcription has been initiated, the cyclin-dependent kinase phosphorylates the carboxy-terminal domain of the RNA Polymerase II complex, triggering the transition from the initiation phase to the elongation phase. The RNA Polymerase II complex then begins to traverse the DNA strand while the TFIID complex remains anchored at the TSS [12].

The mediator complex was first biochemically purified in the late 1990s, but its importance to transcriptional regulation has only become clear in the mid 2000s [14]. It is a large and variable complex containing about 20 subunits and integrates regulatory signals from specific TFs into the transcriptional machinery [5,14]. It has been shown to be essential for the transcription of most genes in yeast and *C. elegans* because it bridges the gap between the RNA Polymerase II complex at the TSS and the specific TFs, which may be bound to the DNA sequence several hundred (or thousand) bases away [14,15]. Mediator contains at least three subcomplexes or modules, including the head, middle and tail modules. The head and middle modules interact with the RNA Polymerase II and TFIID complexes while the tail module interacts with the specific TFs bound further upstream, causing the DNA sequence between them to form a loop [15,16]. Mediator does not bind DNA directly and participates only in protein-protein interactions; once transcription is initiated, the complex dissociates from the progressing RNA Polymerase II complex and remains at the TSS. There are several other interchangeable modules and subunits that are only present in the complex under certain circumstances and impact the binding strength and specificity of the interactions [15]. Mediator has been shown to be particularly important for the activity of nuclear hormone receptors such as thyroid hormone, sterol hormone, and vitamin D receptors. In the presence of the appropriate hormone, the receptors enable the mediator complex to recruit coactivators, acetylate histones, and activate transcription of specific genes, while in the absence of the hormone, the combined complex recruits corepressors, deacetylates histones and maintains the chromatin of those same genes in an inactive state [5,14,15]. As is the case for the other complexes and transcriptional regulatory complexes, the mediator complex is under active research.

When all of the necessary elements are in place at the gene promoter, including the specific TFs, the mediator complex, the general TFs, and the RNA Polymerase II complex, transcription of the gene becomes possible. Transcription of the RNA molecule occurs in three phases: initiation, elongation, and termination [17]. During the initiation phase, the RNA Polymerase II enzyme and associated cofactors bind to the TSS and begin to synthesize the first few bases of RNA [17]. Binding of the RNA Polymerase II complex to the TFIID and mediator complexes results in the phosphorylation of the carboxy-terminal domain which produces a conformational change in the RNA Polymerase II enzyme that subsequently prompts the transition to the elongation phase [14]. During the elongation phase, the RNA Polymerase II enzyme slides rapidly along the DNA strand and synthesizes the rest of the transcript. Chromatin remodelling subunits remove the histones from the DNA in front of the polymerase enzyme and reattach them to the DNA behind [18]. During the first transcriptional event of a gene, subunits of the enzyme may also acetylate the histones attached to the DNA of the gene to increase the efficiency of generation of further transcripts [17]. Once the RNA Polymerase II enzyme reaches the end of the gene, it dissociates from both the DNA strand and the nascent RNA strand and the carboxy-terminal domain is dephosphorylated [5]. The DNA strand may form a large loop structure, bringing the transcriptional termination site close to the TSS so that the enzyme can be recruited to the promoter of the same gene and produce further transcripts [17].

The order of the steps, including formation and binding of the three complexes, and transcriptional initiation and elongation is not entirely clear and may vary between genes. At first glance, it may seem logical that the specific TFs would bind first, they would then recruit the TFIID complex, which would in turn recruit the RNA Polymerase II and mediator complexes, and transcriptional initiation would not proceed until after all three complexes were present. However, it has been shown that the rate-limiting step of the entire procedure is the transition from initiation to elongation [17]. The complexes may combine at the promoter in any order, and in fact many eukaryotic promoters have an RNA Polymerase II enzyme and TFIID complex associated with them even though the gene is not being expressed. The transition to elongation is triggered by the phosphorylation of the carboxy-terminal domain, which only occurs once all of the complexes are in place [14,17]. The RNA Polymerase II enzyme may also initiate synthesis of the first few bases of the RNA transcript and then pause or abort the transcript if the specific TFs and mediator are not present. This allows gene expression to be switched on rapidly once the specific TFs and the mediator complex are bound [14,17].

As the RNA is being synthesized, it is modified in three important ways: a modified guanine is attached to the 5’ end (the 5’ cap); internal sections of the RNA strand are removed and the remaining pieces spliced together; and after transcription is complete the RNA strand is cleaved at a specific site and a string of adenosines is added onto the resulting 3’ end [6]. All three of these operations are performed by a variety of subunits and enzymes that are attached to the phosphorylated carboxy-terminal domain of the RNA Polymerase II complex [6]. Capping is associated with the transition from initiation to elongation [17] and improves mRNA stability by protecting it from specific RNA exonucleases [6]. Splicing is performed by a mini-complex of small nuclear RNAs and proteins called the spliceosome [6]. During each splicing procedure, the small nuclear RNAs recognize the borders of the intron that is to be spliced out while the exons are coated with serine and arginine-rich proteins. Splicing regulates translation of the subsequent mature mRNA in two important ways. The serine and arginine-rich proteins improve the efficiency of translation with the result that spliced mRNAs produce more protein per transcript than unspliced mRNAs [17]. Splicing also impacts the translated product directly: the exclusion of an essential coding exon or the inclusion of a stop codon-containing "poison exon" can prevent the transcript from being translated correctly [17]. As the RNA Polymerase II proceeds past the polyadenylation signal, the RNA is cleaved and polyadenylated, and the RNA Polymerase II complex dissociates from the DNA. The processed mRNA is packaged with RNA binding proteins to sequester the newly formed transcript from further interactions with the DNA strand and thereby prevent DNA damage. Regions of DNA that are under active transcription are often associated with a nuclear pore complex so that the transcript can be transported to the cytoplasm as soon as transcription and processing are completed [17]. If any of the processing steps fail, the transcript will be bound by proteins that mark it for immediate degradation and a nuclear exosome is recruited to degrade the transcript. This avoids the production of malformed or nonsense-containing proteins which may otherwise have deleterious effects.

There are two related aspects of transcription and translation in nematodes that have not been observed in the nuclear genomes of insects or vertebrates: trans-splicing and operons. Seventy percent of *C. elegans* transcripts are trans-spliced during transcription: the original 5' end of the transcript is removed and replaced with one of two 22 bp sequences that originate from small nuclear RNAs. The trans-splice acceptor site on the mRNA has a canonical sequence of UUUCAG and is typically located about 20 bp upstream of the translation start site. The purpose of trans-splicing is unknown but it is suspected to play a role in the initiation of translation [19]. A side-effect of trans-splicing is that it makes it impossible to determine the TSS from cDNA sequence because some or all of the 5' untranslated region is lost during the trans-splice procedure. Trans-splicing takes the place of the 5' capping step (the donated sequence is already capped) and makes it possible for the *C. elegans* genome to organize its genes into operons.

Operons are clusters of genes, close together on the same strand, that are all transcribed after a single RNA Polymerase II initiation event. As the first gene is transcribed, it is processed in the regular way (trans-spliced, cis-spliced, and polyadenylated). After the RNA Polymerase II reaches the end of the first gene, instead of terminating, it continues to transcribe the other genes, which are each trans-spliced and processed in turn. The polyadenylation and trans-splice acceptor sites cause the single transcript to be cleaved into separate mature mRNAs for each gene. Genes in operons are often coexpressed, but due to different rates of mRNA degradation may not always display correlated expression patterns [19].

1.1.3 Post-transcriptional Regulation

Once the gene has been transcribed, it is subject to further regulation at the RNA level. A large quantity of noncoding RNA is transcribed from defined noncoding RNA genes. Many of these ncRNA genes are clearly understood, such as ribosomal RNAs, transfer RNAs, and small nuclear RNAs. However, more recent research has shown that other classes of ncRNAs, including microRNAs and short interfering RNAs, are directly involved in the regulation of transcription, translation, and degradation of protein-coding mRNAs. The *C. elegans* genome contains at least 1300 noncoding RNA genes including at least 120 microRNA genes [20]. MicroRNAs are typically 20 to 22 nucleotides long and tend to be partially complementary to the 3' untranslated region of the mRNAs that they regulate. When they are coexpressed with a target gene, they form a double-stranded RNA complex that is immediately degraded, preventing translation of the target gene [21]. Several *C. elegans* microRNA genes, such as *lin-4* and *let-7*, have been shown to be conserved in other eukaryotes including mammals, indicating that this is an ancient and efficent mechanism of gene regulation [20].

Short interfering RNAs are 21 to 25 nucleotides long, and like microRNAs, block translation of mRNAs by forming a double-stranded complex [21]. Unlike microRNAs, short interfering RNAs are not involved in the regulation of ordinary genes but instead promote the degradation of transcripts from viruses and retrotransposons [22]. For both microRNAs and short interfering RNAs, the double-stranded complex is degraded by the RNA induced silencing complex, subunits of which have been found in all eukaryotes [22].

In addition to microRNA genes and short interfering RNA genes, more poorly defined noncoding RNAs have been detected that are transcribed from intergenic regions and introns of protein-coding genes [21]. Transcription of these RNAs has been shown to interfere with the binding of TFs and RNA Polymerase II to the DNA, thereby preventing the transcripton of a nearby gene [21]. All three types of regulatory noncoding RNAs have been detected in *C. elegans*, but except in a few well-defined cases, their impact on gene regulation is poorly understood.

1.1.4 Focus

For the purposes of this research we have focused primarily on specific TFs and their binding sites in the DNA near the TSS, and also focused on protein-coding genes rather than RNA genes. We did not investigate the role of pre-transcriptional and post-transcriptional regulation mechanisms, only co-transcriptional regulation characterized by the binding of specific TFs to evolutionarily conserved sites in the immediate upstream region of the genes. The purpose of this research is to determine which portions of the upstream sequence are specific transcription factor binding sites (TFBSs).

Terminology note: we will use TFBS to refer to a DNA sequence that is bound by a TF. A regulatory element is a DNA sequence that is involved in the regulation of a nearby gene; if the sequence is removed or altered, the expression of the gene is affected. Sometimes the terms TFBS and regulatory element are used interchangeably (both here and in the literature) under the assumption that a regulatory element functions via the binding of a TF. An enhancer is usually defined as a cluster of TFBSs which are far from the gene with which they interact, but the term may also be used interchangeably with regulatory element. Enhancers are usually discussed with respect to *Drosophila* and vertebrate genomes because TFBSs have not been shown to occur in distant clusters in *C. elegans*.

1.2 Laboratory Investigation of Gene Regulation and Expression

There are a wide variety of laboratory techniques with which to investigate gene regulation. Most methods are gene-centred: they investigate regulatory sequences upstream of one or more genes. Other methods are TF-centred: they investigate the targets of a specific TF [23]. Additionally, because regulatory elements tend to be conserved among coexpressed and orthologous genes, identification of such gene sets can greatly aid the discovery of new TFBSs.

1.2.1 Gene-centred Approaches

Because it can be difficult to accurately detect the expression level of a gene directly, a powerful technique to observe and estimate the timing and location of gene expression is through a reporter gene construct. *Caenorhabditis elegans* is a transparent organism, and therefore the most common reporter genes are for fluorescent proteins such as green fluorescent protein (GFP) and other colours such as yellow and red fluorescent proteins [7,24]. In a typical experiment, the upstream region of the gene of interest is amplified by the polymerase chain reaction (PCR) and cloned into a vector containing the GFP coding sequence. A large number of copies of the construct are injected directly into the *C. elegans* hermaphrodite gonad and the GFP is subsequently expressed in the next generation of worms in approximately the same tissue and developmental stages as the gene of interest. The expression of GFP can be observed directly and photographed, or quantified [25]. To improve specificity, nuclear localization signals can be added to the GFP protein sequence to sequester the GFP to the nucleus of cells within which it is expressed. Alternatively, an in-frame translational fusion construct can include the cellular localization signal of the original gene with the GFP to indicate where the gene product is normally localized. The specific sections of each upstream region responsible for the observed expression pattern can be explored through scanning deletion mutagenesis: numerous GFP expression constructs are created for each upstream region, and a short section is deleted for each one. Those constructs for which the expression is lost indicate which specific portions of the upstream region are necessary for the expression of the gene and are therefore putative regulatory elements.

Once a putative regulatory element has been identified, the next logical step is to test the DNA sequence for protein binding using an electrophoretic mobility shift assay. In this assay, the short DNA sequence (labelled with biotin or radioactive phosphorus atoms) is incubated both alone and with an extract of nuclear proteins; both samples are run on a nondenaturing gel. If the DNA binds a protein in the nuclear extract, the gel will show a shifted band in the column containing the DNA sequence with the nuclear extract but not the DNA sequence alone. To show specific binding, the labelled DNA and nuclear extract can further be incubated with the same unlabelled DNA sequence, and with unlabelled varied DNA sequence as competitors. If the protein binding is specific to the DNA sequence, the lanes with the same sequence as a competitor will show a reduction in the shifted band, while the lanes with a slightly different DNA sequence will not show a reduction in the shifted band. To prove the identity of the protein, an antibody specific to the predicted protein can be added as well; if the antibody binds to the protein that is bound to the DNA, gel lanes containing the antibody will show a third, supershifted band [26].

DNase I endonuclease cleaves double-stranded DNA that is not bound by proteins, and hence is useful for two different protein-binding assays. The DNase I footprinting (or protection) assay is used to limit TF binding to specific bases of a specific sequence of DNA (for example a sequence that has already tested positive in an electrophoretic mobility shift assay). Labelled DNA sequence is incubated both alone and with a nuclear extract, and both samples are briefly exposed to DNase I and then run on a sequencing gel. The DNA without nuclear extract will be cleaved at every base, forming a band on the gel for every base of the sequence. In contrast, for the DNA with nuclear extract, those bases of the DNA covered by protein will not be digested by DNase, leaving a gap in the ladder [26].

The DNAse I hypersensitivity assay is used to detect genomic regions that are free of nucleosomes. Very little DNA in a given cell type is not wrapped around nucleosomes; nucleosome-free regions are typically left open to allow binding of TFs and trancriptional complexes, which is why these regions are associated with enhancers and promoter sequences in mammals. In this assay, nuclear DNA is digested with DNase I, and the digested DNA is extracted and mapped back to the genome using a southern blot or sequencing technique. On a genome-wide scale, the DNase I hypersensitivity assay can be used to identify putative regulatory elements, highly conserved regions, and promoters under active transcription [27].

The yeast-1-hybrid method is a high-throughput approach for finding combinations of DNA binding proteins and their binding sites [28]. In this technique, a cDNA expression library is created in which each open reading frame is fused in-frame to a Gal4p activation domain (the "prey"). A second library contains a large set of promoter sequences linked to the coding sequence of a reporter gene (the "bait"). Yeast are transfected with one construct from each library; only those cells that contain a combination of bait and prey where the prey construct expresses a protein that binds to the bait construct will express the reporter gene. A large number of bait and prey combinations can be tested simultaneously, and for cells that test positive, both constructs can be sequenced. The advantage of this method is that it finds a large variety of DNA binding proteins regardless of whether they function as activators or repressors *in vivo*. The disadvantage is that like other yeast-hybrid methods, it is prone to false positives, and results need to be validated using a different method for confirmation of the protein-DNA interaction. The yeast-1-hybrid method has been used with great success to find novel DNA binding proteins in *C. elegans* [29].

1.2.2 Transcription Factor-centred Approaches

SELEX (Systematic Evolution of Ligands by EXponential enrichment) is an *in vitro* method that is used to rapidly determine a series of putative binding sites for a TF [30]. Briefly, protein extracts are incubated with a large variety of double-stranded DNA oligonucleotides that all have the same end sequences. Antibodies are used to precipitate the TF of interest, or if no antibodies are available, the cell line or organism is transfected with an epitope-tagged variation of the TF. All precipitated DNA is amplified by PCR using primers that match the terminal sequences of the oligonucleotides. Several rounds of immunoprecipitation followed by PCR amplification are done, and eventually only sequences that bind the protein with high affinity remain in the sample; these are subsequently cloned and sequenced [30]. Several variations and optimizations of the technique exist to adapt it to RNA binding proteins and proteins for which some binding sites are known [31]. The advantages of the SELEX method are that it generates a high-quality set of TFBSs for a given TF and tends to find many binding site sequence variations. The disadvantages are that it does not produce genomic binding sites that can subsequently be used to investigate the regulation of specific genes, and some proteins may only bind a subset of the sequences *in vivo* that they bind *in vitro*.

Chromatin immunoprecipitation (ChIP) is used to detect regions of the genome that are associated with specific proteins, either TFs or histones with specific modifications [32]. In this technique, nuclear DNA is extracted from the tissue of interest, and all proteins attached to the nuclear DNA are crosslinked to the DNA using formaldehyde. The DNA is then sheared into small pieces, and antibodies are used to precipitate the protein of interest and all DNA associated with that protein. After this the crosslinks are reversed and the proteins are degraded, leaving only DNA that is highly concentrated for subsequences that are bound by the protein. The advantage of ChIP is that it finds only sites that are bound by protein *in vivo*. The disadvantages are that the method requires a specific antibody to the DNA binding protein, and that it finds all DNA near the protein, even if it is not bound directly, and the precipitation step is not completely accurate. The result is that many precipitated sequences do not contain a binding site for the TF, but it is not clear which are false positives and which are sequences that are bound by a protein that was bound to the TF due to DNA looping [33].

There are a number of subsequent methods following ChIP that can identify the DNA sequences in the sample. The classical method is to use a Southern blot or PCR to determine whether any of the DNA precipitated by the antibody matches one or more candidate sequences, but this method only identifies one sequence at a time [2,33]. More recently, a series of high-throughput methods have been developed to identify most or all precipitated DNA sequences simultaneously. In the ChIP-chip method, the enriched DNA is hybridized to an array containing segments of genomic DNA [34]. Result quality from ChIP-chip is limited by the size of the genome and the number of sequences on the microarray or oligonucleotide array. For large genomes, sometimes arrays are used that contain only promoter sequences or sequences from specific genomic regions, but results from such an array are not genome-wide [2]. In the ChIP-PET (paired-end tags) method, the DNA fragments are cloned and each end is sequenced [35]. The paired-end tags are then mapped back to the genome – a computationally complex task. Similarly, in the ChIP-SAGE method, the fragments are cloned and a short sequence tag is extracted from them, after which the tags are concatenated and sequenced [2]. These two methods were used in the interim while new, inexpensive sequencing technology was being developed. The most recent method to identify the enriched DNA sequences is called ChIP-Seq [36]. In this method, the first 30 to 100 bp of all precipitated DNA fragments are sequenced using the recently developed high-throughput massively parallel short-read sequencing methods such as the Illumina 1G Genome Analyzer sequencing platforms. The sequences are mapped back to the genome and extended across their estimated length, and regions of the genome containing a large number of overlapping results are identified. This method has been shown to provide deep coverage of relevant binding sites using a small amount of enriched DNA starting material, with the result that it generates fewer artifacts than older sequencing methods [33]. ChIP-Seq has successfully been used to find binding sites for both TFs such as STAT1 [36,37] and histones with specific methylations, acetylations, and other variations [38,39].

1.2.3 Gene Expression and Orthologue Identification

Accurate and timely expression of genes is important for the survival of an organism, and as a result we have two important expectations with respect to TFBSs. First, we expect that genes that are expressed at the same time or in the same tissue may be regulated by the same TF and therefore may contain similar binding sites in their upstream regions. Secondly, we expect TFBSs that deliver essential or advantageous gene expression patterns for an organism to be conserved through evolution while DNA sequence that does not impact gene expression will not be under evolutionary constraint.

One way to discover new TFBSs is to look for similar sequence patterns in the upstream regions of coexpressed genes: genes that are all transcribed in the same tissue at the same time. In order to find sets of coexpressed genes, it is necessary to measure the expression of all or most genes simultaneously. Serial analysis of gene expression (SAGE) is a quantitative measure of mRNA levels in a tissue sample [40]. In the SAGE method, an elegant series of biochemical steps is used to extract a short representative sequence from each mRNA (Figure 1.2). The occurrences of each tag are counted, and the resulting tag counts are analogous to the quantity of mRNA transcripts from each gene in the original sample. A list of tags and their respective counts from a particular tissue is called a SAGE library. SAGE has been used to measure gene expression in a wide variety of tissues, notably human cancer tissues [41]. For *C. elegans*, the BC *C. elegans* Gene Expression Consortium has produced 31 SAGE libraries from various *C. elegans* tissues [42]. Two pairs of SAGE libraries are analyzed in Chapter 2.

Another fruitful location in which to search for TFBSs is in the upstream regions of orthologous genes from different species. Orthologues are genes from different species that have very similar coding sequence because they originate from the same gene (through speciation events, not duplication) in the most recent common ancestor of the two species [43]. Orthologue assignments are made by comparing the protein or DNA sequences of all genes in one species to the sequences of all genes in another species. Orthologous genes will be more similar to each other than they are to any other genes in the genome of the other species. However, species-specific paralogous expansions may confound orthologue assignment methods. If the genome of one of the species contains a gene duplication that occurred after the two species diverged, both versions of the gene may be equally similar to the single gene in the other species in terms of sequence similarity. One of the two orthologues may be more similar than the other in terms of function, but it is not possible to tell which one this is from sequence analysis alone. For gene regulation analysis, one-to-many orthologue assignments should be avoided and only one-to-one orthologue assignments should be used because we only expect regulatory elements to be conserved for genes that perform the same function.

Inparanoid is a web resource containing orthologue assignments based on reciprocal best BLAST alignments [43]. Comparisons and orthologue assignments from published genomes are posted at the Inparanoid website [44]. Another method to generate orthologue predictions is with the WABA alignment algorithm, which has been optimized to find long matches in one genome for protein-coding sequences from another genome [45]. WABA finds orthologues for a protein-coding DNA sequence without requiring pre-existing gene predictions by searching for alignments where the first two bases of each amino acid codon are weighted more heavily than the third base in the codon, because for most codons the third base can vary freely without impacting the amino acid sequence. The results from WABA and Inparanoid are compared in Chapter 3.

1.3 Transcription Factors

There are such a large number of proteins that are involved in transcription at every stage that the term "transcription factor" is not well-defined. In this thesis, when we refer to TFs, we are mainly discussing specific TFs that bind specific DNA sequences and impact the expression of a subset of genes under specific circumstances rather than general TFs that are involved in the transcription of most or all genes. TFs contain a transcription regulation domain that interacts with other proteins and cofactors and a DNA binding domain (DBD) that interacts with DNA [11]. They frequently function as homo- or heterodimers or bind DNA in a complementary way [46]. Some TFs are alternatively spliced; different isoforms may have both different protein-DNA interactions and different protein-protein interactions (such as dimerization) [46]. Variations such as isoforms and dimerization mean that the number of different ways TFs can impact gene regulation is potentially very large [11,46].

The transcription regulation domains are generally domains that form strong protein-protein interactions, and may either activate or repress transcription, or both, depending on context [23]. These domains often depend on the binding of coactivator proteins and non-protein ligands, or may need to be phosphorylated in order to function as activators [11]. Many TFs are shuttled between the cytoplasm and the nucleus while undergoing such modifications and interactions.

TFs are usually classified based on the secondary structure of the DBD, and there are more than one hundred different families and subfamilies of these domains [11,23]. DBD families evolve more slowly than other protein sequences and are conserved across many species; multiple examples of orthologous TFs between *C. elegans* and humans have been found [46]. The DBDs from different genes are sometimes very similar to each other and may bind the same sequences and regulate some of the same genes or compete with each other for the same binding sites [46]. Most types of DBDs interact with the DNA via an alpha helix that is positioned in the major groove of the DNA strand; amino acid residues in the alpha helix form hydrogen bonds and van der Waals interactions with the bases of the DNA strand [11]. A few of the most common types of DBDs are described below.

Homeodomain TFs contain a highly conserved 60 amino acid domain and are found in many species including *C. elegans*, *Drosophila*, and humans [11,46]. They tend to have important functions such as control over general body organization [11]. They are part of a larger superfamily of DBDs called the helix-turn-helix structural motif that is characterized by two alpha helices (one of which interacts with the major groove of the DNA) that are connected by a short chain of amino acid residues [26]. Another subset of the helix-turn-helix superfamily is the winged helix domain, which has an additional alpha helix and a beta sheet in addition to the helix-turn-helix structural motif. Winged helix TFs often contain two DBDs which bind a palindromic DNA sequence in a symmetrical fashion [47].

The helix-loop-helix DBD contains two alpha helices packed closely together and linked by a short loop. Like the helix-turn-helix superfamily, there are many variation of this structual motif [11]. TFs containing this domain tend to bind DNA as homo- or heterodimers [26].

Zinc finger domains have a protein structure in which a zinc 2+ ion is coordinated by four amino acid residues such as cysteine and histidine; there are several variations [11]. Zinc finger-containing TFs are very common in most species and the zinc finger domains tend to occur in clusters with multiple fingers interacting with a continuous stretch of DNA [26]. Another type of DBD that bind zinc are the nuclear hormone receptors [11]. These TFs interact with lipophilic hormones that freely diffuse through the cell membrane, such as steroid hormones and vitamin D; their activation domain is only functional when they are bound to their hormone ligand and bind as homodimers [11]. This type of TF is extremely common in *C. elegans*; *C. elegans* contains an estimated 274 different nuclear hormone receptors while humans have only 43 [46].

Leucine zipper or bZip DBDs are characterized by a long alpha helix that has a row of leucine (or other hydrophobic) residues along one side of the helix at one end only. These TFs form homo- and more often heterodimers through van der Waals interactions between the leucine residues on each subunit [26]. The other end of the alpha helix contains basic amino acid residues that interact with the major groove of the DNA strand [11].

For species with sequenced genomes, most TFs can be predicted based on sequence similarity to known DBDs. However, not all genes containing predicted DBDs are true TFs. For example, some genes with zinc finger domains may be RNA binding proteins or may only bind other proteins. Recently, two curated lists of predicted *C. elegans* TFs have been published which estimated the total number of TFs at 934 and 664 respectively [7,46]. In one study, 127 different families of DBDs were represented [7], and in the other study, 23 *C. elegans* TFs were found to contain two DBDs from different families [46]. It is expected that binding sites for human TFs can be used to predict binding sites for *C. elegans* TFs and vice versa [46].

1.3.1 Transcription Factor Binding Sites

TFs can usually bind to a number of different sequences, with a stronger affinity for some sequences than others. The range of sequences they can bind to depends on the TF concentration [48], and endogenous binding sites for each TF are typically a series of similar but not identical sequences. For example, the *C. elegans* TF DAF-19 contains a winged helix DNA-binding domain (Figure 1.3 A) [49] and has been experimentally shown to bind a variety of similar, partially palindromic sequences that are 14 bp wide (Figure 1.3 B).

A simple way to represent a variety of different sequences is the International Union of Pure and Applied Chemistry (IUPAC) consensus sequence, in which the four DNA letters can be substituted by other letters that represent two or more DNA bases simultaneously (Figure 1.3 C; Table 1.1) [50]. The advantage of the consensus sequence is that the sequence representation is very brief and is displayed in simple text. However, it contains no information as to how often each base appears in each position across orthologues, for example.

A position frequency matrix (PFM) is simply a table containing the sum of the frequencies of each base in each position from a set of TFBSs (Figure 1.3 D). A PFM can be normalized to make the sum of each column (or the sum of the entire table) equal to one. The advantage of the PFM is that it summarizes and digitizes a list of similar sequences, and can be used to assign a similarity score to other sequences that are suspected to bind the same TF.

However, it is not a complete representation of the binding sequence set: the PFM loses information about which specific combinations of bases occur.

A sequence logo is a multicoloured visual representation of a PFM (Figure 1.3 E) [51]. Along the X-axis of the logo are the positions of the binding sequences just as in the PFM. The Y-axis shows the information content (IC) measured in binary digits (bits) at each position. A perfectly conserved base has an IC of two bits just as a single DNA sequence has one of four bases, and therefore two bits of information, in each position of the strand. A position that is filled by one base in half of the examples and a second base in the other half of the examples has an IC of one in total. The error bars are an indication of how many sequences were used to generate the sequence logo: the more sequences that were used, the less it would be impacted by the addition or removal of one sequence and therefore the smaller the error bars.

Three databases of experimentally validated TFBSs were used in this research, ORegAnno [52], TRANSFAC [53] and JASPAR [54]. ORegAnno was created at the Genome Sciences Centre by my colleagues Dr. Obi Griffith and Dr. Stephen Montgomery and consists of TFBSs that were curated from peer-reviewed publications and recorded in their genomic context. The record of genomic context in ORegAnno was a key improvement over other TFBS databases and the ORegAnno sites were used throughout this research as positive controls for TFBS prediction. ORegAnno currently contains 192 *C. elegans* TFBSs and 14 166 TFBSs in other species.

TRANSFAC is a relational database that contains 319 TF binding sequences and PFMs. The TRANSFAC data is derived from published experiments, and the focus of the database is on eukaryotic gene regulation. JASPAR is an open-access, highly curated and non-redundant database of 59 TFBSs determined mainly by *in vitro* DNA binding experiments such as SELEX. The usefulness of both TRANSFAC and JASPAR to this research is limited to TF binding models; neither can provide positive controls for genomic TFBS prediction because none of the sites are tied to genomic locations that were experimentally shown to bind the TFs *in vivo*.

1.4 Nematodes

This research was carried out in the model organism *C. elegans* and other nematodes. Phylum Nematoda consists of a diverse collection of invertebrates, ranging in length from 1 mm (*C. elegans*) to 17 cm long or more (*Ascaris* intestinal roundworms) [55]. Non-parasitic nematodes subsist in a wide variety of natural habitats, including both dry and moist soil, seawater, arctic ice, and sulfur-rich sediments [56]. Parasitism has evolved independently in phylum Nematoda on at least four different occasions [57]. Parasitic nematodes are responsible for a variety of human diseases including hookworm, river blindness, and guineaworm disease, as well as diseases of plants (including food crops such as potatoes and soybeans) and animals (both domestic and wild) (Figure 1.4).

Most species of nematodes are gonochoristic; that is, they consist of both females and males [58]. However, various lineages of nematodes have independently evolved a hermaphrodite/male system of reproduction [59]. In these species, the hermaphrodites are essentially females that produce sperm for a brief period during early sexual development. They store the sperm, complete their development of female reproductive organs, and then use the sperm to fertilize their own eggs. This means that a single hermaphrodite can populate a new territory by itself. Males still exist in these species, and can fertilize the hermaphrodites' eggs, but they only consist of 0.02% of the wild population [60].

C. elegans was described as a species in 1901 [61] and was suggested as a model organism by Dr. Sydney Brenner in the late 1960s [62]. *C. elegans* is a free-living bacteriovore that reproduces by the hermaphrodite/male system. It features an invariant, fully characterized lineage of embryonic and developmental cell division [63] and has been shown to be a tractable model for human cellular processes. An example of this is the research of Dr. R. H. Horvitz on apoptosis in *C. elegans* [64], a process which has since been shown to be important to cancer research.

C. elegans was the first multicellular organism to have its genome sequenced: a draft was completed in 1998 [65], and the genome is updated on a monthly basis with gene predictions, annotations, and corrections [66]. The genome of *C. elegans* consists of six chromosomes, including five autosomes (labelled with Roman numerals I through V) and a sex chromosome (labelled X). The genome is only 100 Mbp in total, 1/30th the size of the human genome, but it contains more than 20 000 genes, about 2/3 as many as in the human genome. This makes it especially suitable for studying gene regulation because its genes are close together, leaving much less intergenic sequence to search through for TFBSs. The *C. elegans* genome has been predicted to contain 660 to 900 TFs, of which about 200 are orthologues of human TFs [7,46,67].

In addition to *C. elegans*, the genomes of seven other species of nematodes have been sequenced. Four of them are in the same genus as *C. elegans*, one is in a different genus but is also a soil bacteriovore, and two are human parasites. The other species in genus *Caenorhabditis* are called *C. briggsae*, *C. remanei*, *C. brenneri* and *C. japonica*. These four species of *Caenorhabditis* are highly similar. All eat bacteria, especially *E. coli* [62] and are found in soils all over the world. A primary difference between them is microhabitat: different species of

Caenorhabditis attach themselves to different host species for the purpose of traveling to a new environment. Specifically, *C. elegans* is associated with slugs, *C. briggsae* is associated with snails, and *C. remanei* is associated with woodlice, while other species in genus *Caenorhabditis* are associated with carrion beetles, fruit flies, and palm weevils [68]. In addition, *C. remanei*, *C. brenneri*, and *C. japonica* are gonochoristic while *C. elegans* and *C. briggsae* are hermaphroditic (with some males). Only a few genes are thought to be involved in the conversion from gonochorism to hermaphroditism [69], but it has a large impact on the genome sequencing effort: the gonochoristic species do not inbreed as readily as the hermaphrodites and as a result their genomes are more varied and much more difficult to assemble [70].

The three more distantly related nematodes discussed here are *Pristionchus pacificus*, *Brugia malayi*, and *Trichinella spiralis*. *P. pacificus*, like *C. elegans* is a free-living soil bacteriovore, and it is associated with scarab and potato beetles [71]. *P. pacificus* is also a self-fertilizing hermaphrodite, but it is hypothesized that the ancestors of *C. elegans*, *C. briggsae*, and *P. pacificus* all evolved hermaphroditism independently because intermediate species such as *C. remanei* and *C. japonica* are gonochoristic [59]. *B. malayi* is a mosquito-borne human parasite that causes lymphatic filariasis and elephantiasis [72], and *T. spiralis* is a parasite of pigs (and carnivores) that can be contracted by humans who have eaten undercooked infected meat [73].

The similarity of these species, especially those in genus *Caenorhabditis*, means that they are expected to share many of the same or very similar genes. A comparison of the *C. elegans* and *C. briggsae* genomes indicated that they have about 12 000 clear orthologues [74]. Several studies have shown that orthologous genes between *C. elegans* and *C. briggsae* also share TFBSs [75,76]. Orthologues between *C. elegans* and the other species were predicted as described in Chapter 3.

1.5 Motif Discovery

The SAGE technique can be used to find a set of coexpressed genes. Comparative genomics analysis can lead to the identification of a set of orthologous genes from different species. In both cases, we can generate a set of genes that may contain similar TFBSs in their otherwise disparate upstream regions. When we use computational methods to predict one or more TFBSs in a set of sequences based on sequence similarity, the prediction is referred to as a motif.

The primary goal of computational motif discovery is to find short sequence variations that are found more often than expected in a sequence set, also referred to as a foreground set. The occurrence frequency of a motif in the foreground set is compared to its frequency in a background set in order to establish the motif significance. It is not useful to find a motif that occurs frequently in the foreground in the absence of a comparison background set; that may result in finding a motif that occurs frequently in every sequence set or is very common in the genome. The background set must be chosen carefully and ideally is substantially larger than the foreground, with sequences that originate from the same general source as the foreground sequences. For best results, the foreground and background sequence sets should only differ in one key way. For example, the foreground sequences could be from genes that are coexpressed while the background sequences are from randomly chosen genes that are not coexpressed. Using the correct background ensures that any motifs that are found are associated with the primary characteristic of the foreground set (e.g. coexpression) and not due to a confounding factor.

In this thesis we will discuss two computational methods to find motifs: Gibbs sampling and word-counting. The width of the expected motif must be specified in advance for both methods. When the widths of the hypothetical TFBSs are not known in advance, the methods can be run repeatedly using several widths and the results can be combined or compared to each other.

Gibbs sampling is a stochastic motif discovery method that produces slightly different results each time it is run even when the initial parameters and sequences are the same. In this method, initial motif locations are chosen on each sequence at random. Over a series of iterations, the motif locations on each of the foreground sequences are shifted slightly and eventually converge onto similar sites. Gibbs sampling is a computationally efficient method that requires no pre-processing steps or large-scale data storage. The implementation of Gibbs sampling used in this research is MotifSampler [77]. MotifSampler uses a high-order Markov model to represent the frequency of each subsequence in the background set and is particularly robust at avoiding irrelevant motifs [78].

Word-counting motif discovery algorithms methodically enumerate all n-tuples of bases found in a set of sequences, and then compile groups of similar frequent tuples in order to form the consensus sequence of a motif. The claim of this method is that it performs an exhaustive search and is therefore guaranteed to find the most frequent putative TFBSs in the set. Two word-counting motif discovery algorithms were used in this research: RSAT [79] and DME [80].

RSAT is a web-based tool that requires several steps: first the background set must be uploaded to the website, which produces a background enumeration file, then the foreground is uploaded together with the background enumeration file. RSAT then produces a list of sequences that are found more often than expected in the foreground with respect to the background frequencies. The relative merits of MotifSampler and RSAT are discussed in Chapter 2.

DME is similar to RSAT but more sophisticated in that it can find sequence variations and complete motifs of a specified information content in addition to individual overrepresented sequences. As a result it takes much longer to run and can only be used efficiently on small sequence sets. DME is discussed further in Chapter 4.

1.6 Thesis Overview

1.6.1 Assumptions

For the purposes of this thesis research, we made the following assumptions:

Transcriptional control is a primary element of gene regulation. Although gene expression and function is affected by factors other than mRNA transcription, such as chromatin compaction, histone methylation, and RNA processing, the focus of this research was on gene transcription as measured through SAGE.

Transcription is controlled by TFBSs. Fine-tuned control of gene transcription has been shown to be mediated by TFs that bind to specific sites in the DNA near the genes being controlled; we had numerous examples of such sites from ORegAnno, and our objective was to search for more TFBSs.

TFBSs are found in the upstream regions of genes in nematodes. Most known TFBSs in *C. elegans* were found within 1500 bp of the translational start site. Although a few examples were found in the literature of sites far further upstream and sites in the introns, only the immediate upstream region was shown to be specifically enriched for TFBSs. Any TFBSs that exist outside of this region were not found.

TFBSs consist of 6 to 14 bp conserved motifs. Almost all known TFBSs in *C. elegans* fell into the length range of 6 to 14 bp, and displayed sequence conservation for some or all of the bases. We searched for other motifs that fit this pattern.

Orthologous and coexpressed genes are likely to contain the same TFBSs. The experiments we performed were based on searching in sets of sequences that were expected to be enriched for functional TFBSs. The upstream regions of both orthologous and coexpressed gene sets were investigated for the presence of conserved motifs; both had been previously shown to contain shared TFBSs in nematodes.

1.6.2 Summary of Thesis

The research was conducted in three phases: the first phase concerned TFBSs shared by coexpressed genes, the second phase concerned TFBSs shared by orthologous genes, and the third phase concerned the regulation of genes with a common function.

In the first phase, described in Chapter 2, I worked with two collaborators to investigate gene regulation in specific *C. elegans* tissues, namely intestine and ASE neurons. My objective was to investigate gene regulation in these tissues using bioinformatic techniques such as motif discovery. We compared two SAGE libraries in each study to assemble lists of coexpressed genes for each tissue. Both collaborators possessed information on which TFs were responsible for regulating some of the genes in the tissue of interest. They had also assembled examples of binding sites for those TFs. For the Intestinal study, I performed motif discovery on a small set of specifically-expressed genes in order to obtain a set of putative TFBSs that were important for intestinal expression. For the ASE neuron study, the collaborator had already obtained a set of putative TFBSs. For both studies, I scanned the upstream regions of various sets of genes that were strongly expressed in the tissues of interest and showed that the level of gene expression in those tissues was related to the likelihood of finding a high-scoring match to the TFBS in the upstream region.

In the second phase, described in Chapter 3, I performed a high-throughput comparative genomics analysis of the upstream regions of *C. elegans* protein-coding genes. For each gene in *C. elegans*, I predicted orthologues in the genomes of seven other nematode species. For those genes that had at least three high-quality orthologues, I combined the upstream region of each *C. elegans* gene with the upstream regions of its orthologues to form an orthologous upstream sequence region set. Motif discovery was performed on the upstream sequence region sets to identify conserved upstream motifs, and these motifs were placed in the cis-regulatory element database (cisRED) [81]. I compared the cisRED motifs to known TFBSs from both *C. elegans* and mammalian species and found that 26% of the cisRED motifs were similar. These annotated motifs are candidates for novel binding sites of characterized *C. elegans* TFs and uncharacterized *C. elegans* TFs that are orthologues of characterized mammalian TFs. Other motifs were identified as unannotated protein-coding exons and ncRNA genes.

In the third phase, described in Chapter 4, I compared cisRED motifs to each other and placed them into groups based on sequence similarity. Many of the motif groups were associated with genes that also had functional similarity. I identified a total of 15 motif groups that were specifically associated with genes encoding ribosomal proteins. Eight of the motif groups were extensions and variations of the canonical trans-splice acceptor site. One of the fifteen was tested for regulatory function in a series of GFP expression experiments and was shown to be associated with gene expression in the pharynx.

1.7 Chapter 1 Table

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

Table 1.1 – IUPAC Letter Symbols for DNA Consensus Sequences

Transcription factors frequently bind to a variety of different but related sequences. These letters are used to represent ambiguous DNA sequence [50].

1.8 Chapter 1 Figures

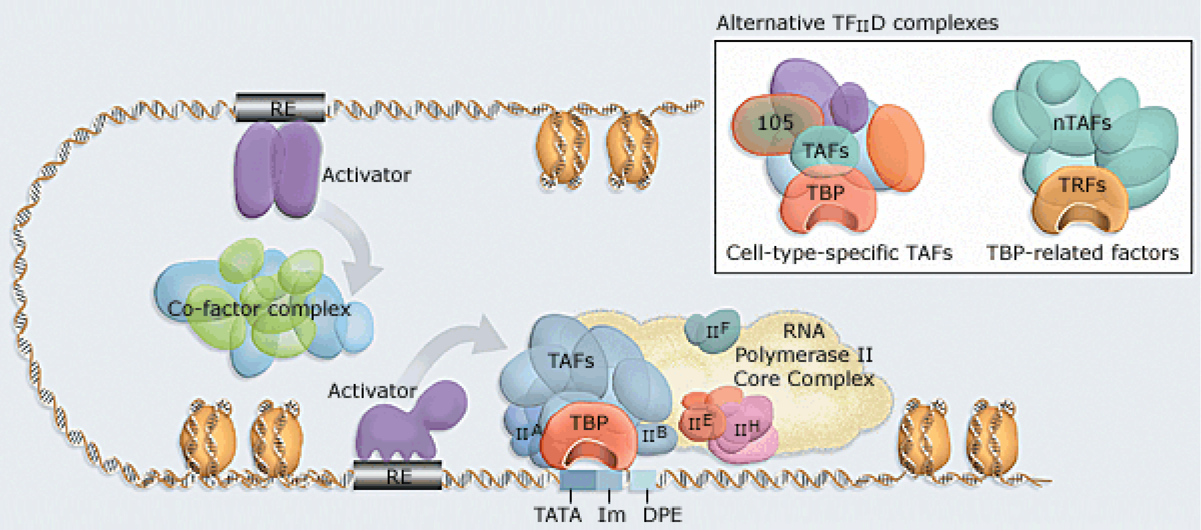


Figure 1.1 – Diagram of Cotranscriptional Gene Regulation

RNA polymerase, which copies DNA into the RNA that subsequently directs protein synthesis, is guided in its work by transcription factors. Some of these factors form multisubunit complexes (cofactors) that serve as bridges between activators, which regulate the rate of transcription, and the RNA polymerase machinery. One class of cofactors, called TAFs, join with TATA binding protein to form the TFIID complex, and attach to the TATA box DNA at the gene's promoter. All cells use an elaborate transcription apparatus to express genes, but some specialized cells (e.g., ovaries, testes, neurons) use alternative versions. For example, Tjian and colleagues have shown that TAFII 105 (box) is specifically required for oocyte formation. Image and caption © 2009 Robert Tjian, Howard Hughes Medical Institute. Used with permission.

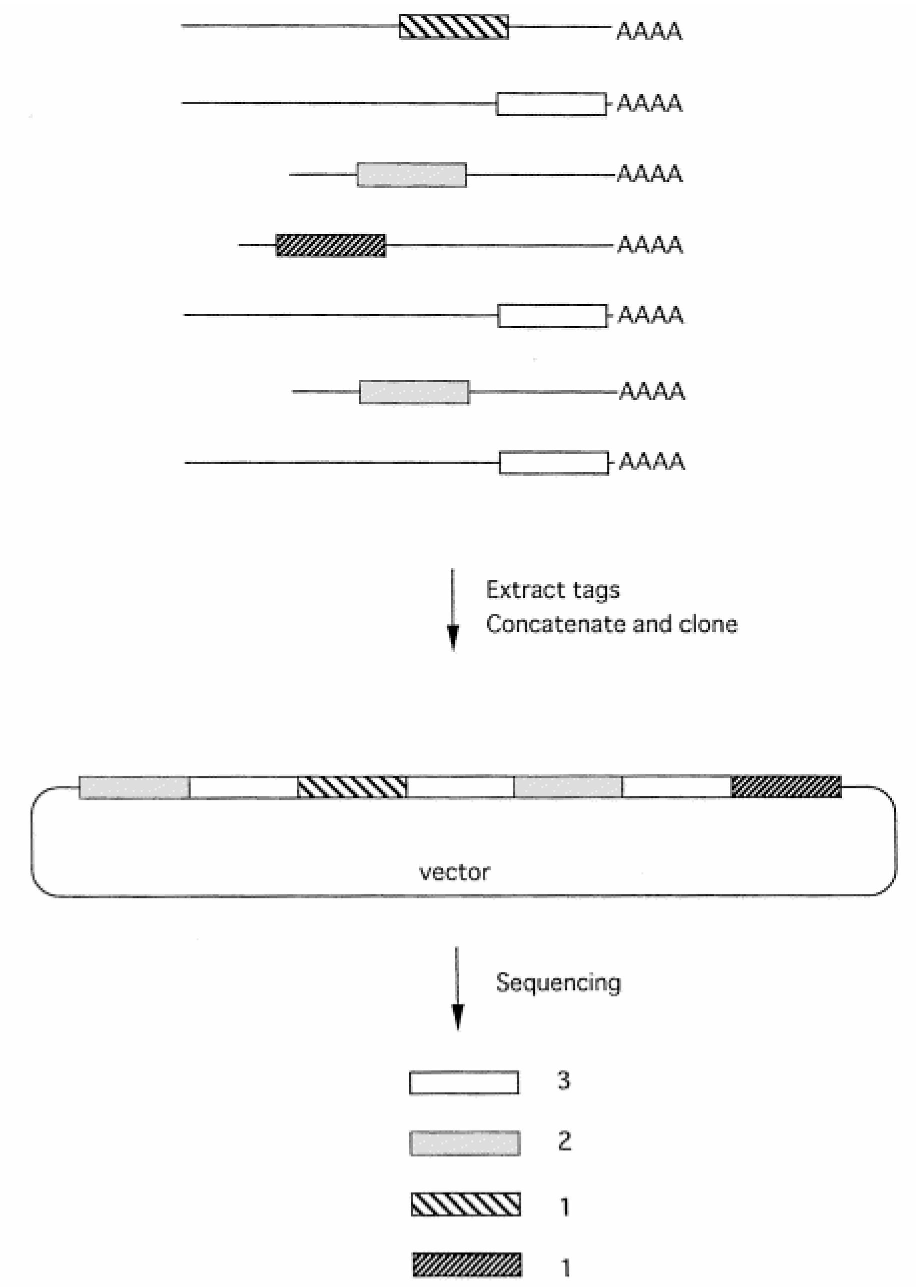
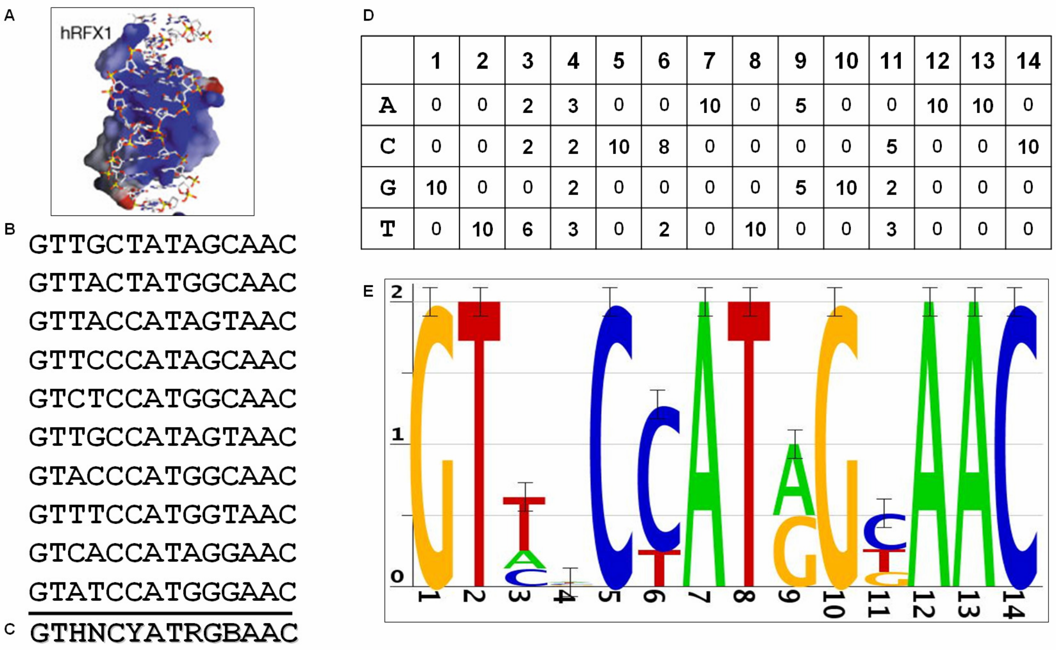


Figure 1.2 – Schematic of SAGE Technique for Gene Expression Quantification

Polyadenylated RNA is extracted from the tissue and converted to complementary DNA for stability. Using a series of restriction enzyme digests, a 14 or 21 bp representative portion (called a “tag”) is extracted from each complementary DNA sequence. Tags are concatenated, sequenced, and counted. Tag sequences are compared back to gene sequences to determine from which gene they originated. Counts of each tag occurrence represent the quantity of transcribed mRNA from each gene that was present in the original tissue. Image reprinted from the Journal of Immunological Methods, Volume 250, by Yamamoto M, Wakasuti T, Hada A, and Ryo A, “Use of serial analysis of gene expression (SAGE) technology”, pp 45-66, Copyright 2001, with permission from Elsevier [82].



GTTGCTATAGCAAC
GTTACTATGGCAAC
GTTACCATAGTAAC
GTTCCCATAGCAAC
GTCTCCATGGCAAC
GTTGCCATAGTAAC
GTACCCATGGCAAC
GTTTCCATGGTAAC
GTCACCATAGGAAC
GTATCCATGGGAAC

GTHNCYATRGBAAC

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	0	2	3	0	0	10	0	5	0	0	10	10	0
C	0	0	2	2	10	8	0	0	0	0	5	0	0	10
G	10	0	0	2	0	0	0	0	5	10	2	0	0	0
T	0	10	6	3	0	2	0	10	0	0	3	0	0	0

Figure 1.3 – Examples of TFBS Representations

A – Three-dimensional structure of the winged helix DNA binding domain bound to its site. Image reprinted by permission from Macmillan Publishers Ltd: in Nature, by Gajiwala KS, Chen H, Cornille R, Roques BP, Reith W, Mach B, and Burley SK, "Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding", Volume 403, pp 916-921, Copyright 2000 [47]. B – Selection of sequences bound by DAF-19, a winged helix domain-containing TF, published by Efimenko *et al.* [49]. C – IUPAC consensus sequence of the TFBS in B (see Table 1.1 for IUPAC symbols). D – Position frequency matrix of the TFBS in B, showing the frequency of each base in each position. E – Sequence logo of the TFBS in B. The Y-axis indicates the information content (IC). Note that a perfectly conserved position has an IC of two (e.g. position 1), a position equally likely to contain either of two bases has an IC of one (e.g. position 9), and a position where all four bases are equally likely has an IC of zero (e.g. position 4).

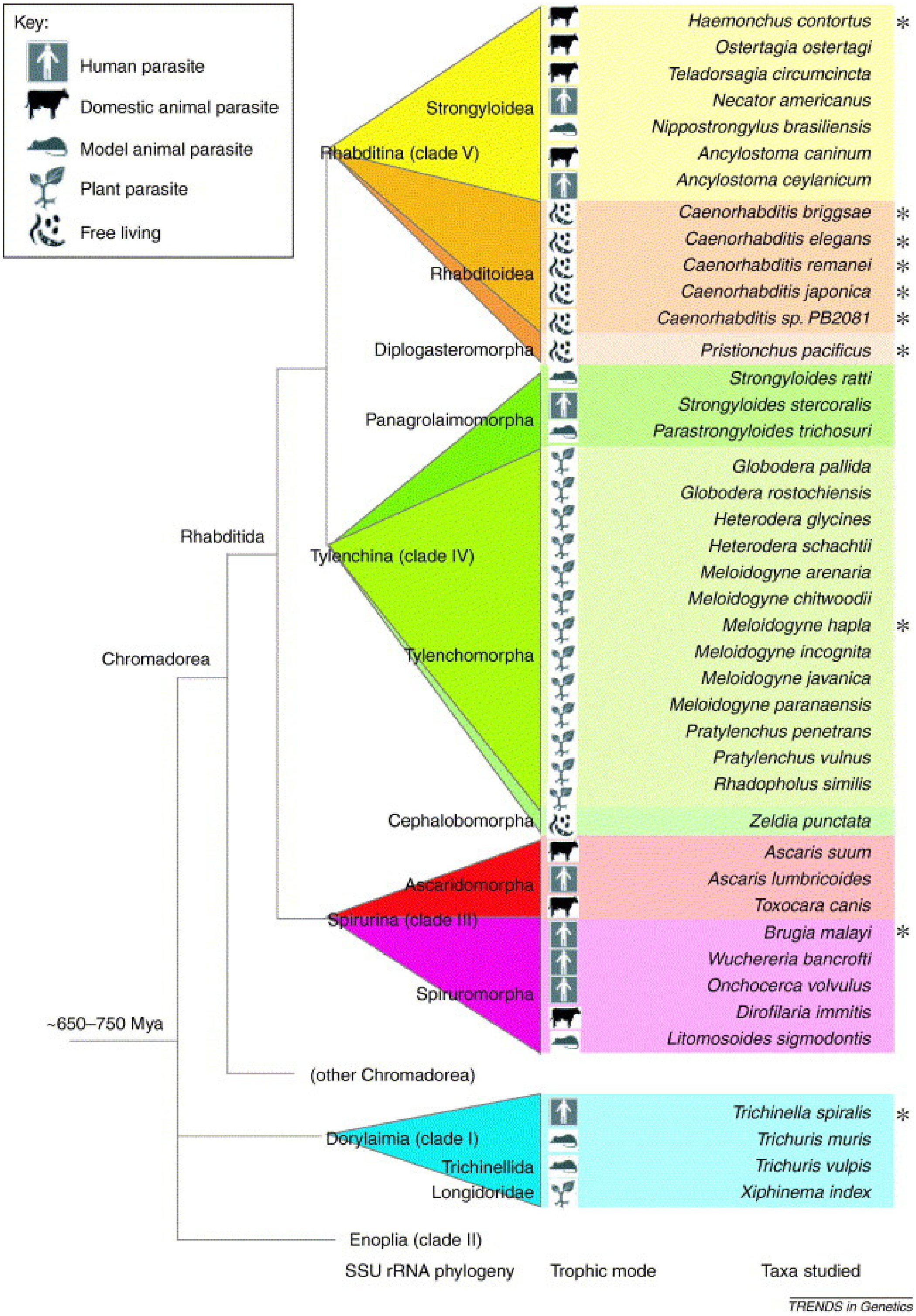


Figure 1.4 – Phylogeny of Phylum Nematoda

The phylogeny of phylum Nematoda is still under active reorganization. Shown is a list of nematode species from which small subunit RNA has been obtained to form a preliminary sequence-based (rather than morphology-based) classification, published by Mitreva *et al.* [57]. Species indicated with asterisks have genome sequence projects completed or underway – all of these except *Haemonchus contortus* and *Meloidogyne hapla* were analyzed as described in Chapter 3. *Caenorhabditis sp. PB2801* was formally named *C. brenneri* after the publication of this figure. Image reprinted from Trends in Genetics, Volume 21, by Mitreva M, Blaxter ML, Bird DM, and McCarter JP, “Comparative genomics of nematodes”, pp 573-581, Copyright 2005, with permission from Elsevier.

1.9 References

1. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36: D475-9.
2. Higgs DR, Vernimmen D, Hughes J, Gibbons R (2007) Using genomics to study how chromatin influences gene expression. Annu Rev Genomics Hum Genet 8: 299-325.
3. Cui M and Han M (2007) Roles of chromatin factors in C. elegans development. WormBook 1-16.
4. Suzuki MM, Kerr AR, De Sousa D, Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res 17: 625-631.
5. Blackwell TK and Walker AK (2006) Transcription mechanisms. WormBook 1-16.
6. Howe KJ (2002) RNA polymerase II conducts a symphony of pre-mRNA processing activities. Biochim Biophys Acta 1577: 308-24.
7. Okkema PG and Krause M (2005) Transcriptional regulation. WormBook 1-40.
8. Gaudet J and Mango SE (2002) Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science 295: 821-825.
9. McGhee JD (2007) The C. elegans intestine. WormBook 1-36.
10. Wenick AS and Hobert O (2004) Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. Dev Cell 6: 757-70.
11. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, et al. (2000) Molecular cell biology. New York, NY, USA: W. H. Freeman and Company. : 1344 p.
12. Cler E, Papai G, Schultz P, Davidson I (2009) Recent advances in understanding the structure and function of general transcription factor TFIID. Cell Mol Life Sci 66: 2123-2134.
13. Cramer P (2004) RNA polymerase II structure: from core to functional complexes. Curr Opin Genet Dev 14: 218-226.
14. Kornberg RD (2005) Mediator and the mechanism of transcriptional activation. Trends Biochem Sci 30: 235-239.
15. Casamassimi A and Napoli C (2007) Mediator complexes and eukaryotic transcription regulation: an overview. Biochimie 89: 1439-1446.
16. Miele A and Dekker J (2008) Long-range chromosomal interactions and gene regulation. Mol Biosyst 4: 1046-1057.
17. Moore MJ and Proudfoot NJ (2009) Pre-mRNA processing reaches back to transcription and ahead to translation. Cell 136: 688-700.
18. Workman JL (2006) Nucleosome displacement in transcription. Genes Dev 20: 2009-2017.
19. Blumenthal T (2005) Trans-splicing and operons. WormBook 1-9.
20. Stricklin SL, Griffiths-Jones S, Eddy SR (2005) C. elegans noncoding RNA genes. WormBook 1-7.
21. Yazgan O and Krebs JE (2007) Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes. Biochem Cell Biol 85: 484-496.
22. Naqvi AR, Islam MN, Choudhury NR, Haq QM (2009) The fascinating world of RNA interference. Int J Biol Sci 5: 97-117.
23. Walhout AJ (2006) Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Res 16: 1445-1454.
24. Zhang S, Ma C, Chalfie M (2004) Combinatorial marking of cells and organelles with reconstituted fluorescent proteins. Cell 119: 137-144.

25. Cao Z, Wu Y, Curry K, Wu Z, Christen Y *et al.* (2007) Ginkgo biloba extract EGb 761 and Wisconsin Ginseng delay sarcopenia in Caenorhabditis elegans. J Gerontol A Biol Sci Med Sci 62: 1337-1345.
26. Yang VW (1998) Eukaryotic transcription factors: identification, characterization and functions. J Nutr 128: 2045-51.
27. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D *et al.* (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16: 123-131.
28. Wang MM and Reed RR (1993) Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. Nature 364: 121-126.
29. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA *et al.* (2006) A gene-centered C. elegans protein-DNA interaction network. Cell 125: 1193-1205.
30. Pollock R and Treisman R (1990) A sensitive method for the determination of protein-DNA binding specificities. Nucleic Acids Res 18: 6197-204.
31. Bouvet P (2009) Identification of Nucleic Acid High-Affinity Binding Sequences of Proteins by SELEX. Methods Mol Biol 543: 139-150.
32. Solomon MJ, Larsen PL, Varshavsky A (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 53: 937-947.
33. Massie CE and Mills IG (2008) ChIPping away at gene regulation. EMBO Rep 9: 337-343.
34. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG *et al.* (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-9.
35. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207-219.
36. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4: 651-657.
37. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221-5231.
38. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE *et al.* (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.
39. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40: 897-903.
40. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487.
41. Strausberg RL, Buetow KH, Greenhut SF, Grouse LH, Schaefer CF (2002) The cancer genome anatomy project: online resources to reveal the molecular signatures of cancer. Cancer Invest 20: 1038-1050.
42. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL *et al.* (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68: 159-169.
43. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041-1052.
44. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res 36: D263-6.
45. Kent WJ and Zahler AM (2000) Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res 10: 1115-25.

46. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA *et al.* (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol 6: R110.
47. Gajiwala KS, Chen H, Cornille F, Roques BP, Reith W *et al.* (2000) Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. Nature 403: 916-921.
48. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Res 13: 2381-2390.
49. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR *et al.* (2005) Analysis of xbx genes in C. elegans. Development 132: 1923-34.
50. Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res 13: 3021-3030.
51. Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.
52. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36: D107-13.
53. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: 108-110.
54. Bryne JC, Valen E, Tang ME, Marstrand T, Winther O *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: 102-106.
55. Garcia-Leiva J, Barreto-Zuniga R, Estradas J, Torre A (2008) Ascaris lumbricoides and iron deficiency anemia. Am J Gastroenterol 103: 1051-1052.
56. De Ley P (2006) A quick tour of nematode diversity and the backbone of nematode phylogeny. WormBook 1-8.
57. Mitreva M, Blaxter ML, Bird DM, McCarter JP (2005) Comparative genomics of nematodes. Trends Genet 21: 573-581.
58. Pires-daSilva A (2007) Evolution of the control of sexual identity in nematodes. Semin Cell Dev Biol 18: 362-370.
59. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F *et al.* (2004) Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A 101: 9003-9008.
60. Zarkower D (2006) Somatic sex determination. WormBook 1-12.
61. Maupas E (1900) Modes et formes de reproduction des nematodes. Archives De Zoologie Experimentale Et Generale 8: 463-624.
62. Brenner S (1974) The genetics of Caenorhabditis elegans. Genetics 77: 71-94.
63. Sulston JE, Schierenberg E, White JG, Thomson JN (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol 100: 64-119.
64. Ellis HM and Horvitz HR (1986) Genetic control of programmed cell death in the nematode C. elegans. Cell 44: 817-29.
65. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012-2018.
66. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C *et al.* (2007) WormBase: new content and better access. Nucleic Acids Res 35: D506-10.
67. Thomas JH (2008) Genome evolution in Caenorhabditis. Brief Funct Genomic Proteomic 7: 211-216.
68. Baird SE (1999) Natural and experimental associations of Caenorhabditis remanei with Trachelipus rathkii and other terrestrial isopods. Nematology 1: 471-475.

69. Haag ES and Kimble J (2000) Regulatory elements required for development of caenorhabditis elegans hermaphrodites are conserved in the tra-2 homologue of C. remanei, a male/female sister species. Genetics 155: 105-16.
70. Barrière A, Yang S, Pekarek E, Thomas C, Haag ES *et al.* (2009) Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res 19: 470-480.
71. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K *et al.* (2008) The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet 40: 1193-1198.
72. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q *et al.* (2007) Draft genome of the filarial nematode parasite Brugia malayi. Science 317: 1756-1760.
73. Takumi K, Teunis P, Fonville M, Vallee I, Boireau P *et al.* (2009) Transmission risk of human trichinellosis. Vet Parasitol 159: 324-327.
74. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR *et al.* (2003) The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. PLoS Biol 1: E45.
75. Gaudet J, Muttumu S, Horner M, Mango SE (2004) Whole-genome analysis of temporal gene expression during foregut development. PLoS Biol 2: e352.
76. GuhaThakurta D, Schriefer LA, Waterston RH, Stormo GD (2004) Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. Genome Res 14: 2457-2468.
77. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9: 447-64.
78. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B *et al.* (2003) Genome-specific higher-order background models to improve motif detection. Trends Microbiol 11: 61-6.
79. Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M *et al.* (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36: W119-27.
80. Smith AD, Sumazin P, Zhang MQ (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc Natl Acad Sci U S A 102: 1560-1565.
81. Robertson G, Bilenky M, Lin K, He A, Yuen W *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res 34: D68-73.
82. Yamamoto M, Wakatsuki T, Hada A, Ryo A (2001) Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods 250: 45-66.

2 Bioinformatic Methods for Investigation of Gene Regulation in Coexpressed Gene Sets[1]

2.1 Introduction

The upstream regions of coexpressed genes often share transcription factor binding sites (TFBSs). Many *C. elegans* researchers specialize in the gene expression and development of a specific tissue type, and require bioinformatic expertise for motif discovery to find the shared TFBSs. Recently, two studies were published regarding the investigation of gene regulation in a specific cell type. One study concerned the regulation of gene expression in the *C. elegans* intestine by the ELT-2 trancription factor (TF) (the "Intestinal study") [1]. The other study concerned the regulation of gene expression in the *C. elegans* ASER sensory neuron by the CHE-1 TF (the "ASE neuron study") [2]. Bioinformatic methods used in both studies were similar in that they involved the identification of coexpressed genes via the comparison of two SAGE libraries. One study featured bioinformatic motif discovery in a small set of specifically expressed genes and their orthologues in *C. briggsae* and *C. remanei*. Both studies featured the use of position frequency matrices to model relevant experimentally validated TFBSs. The primary bioinformatic goal of the studies was to determine how the distribution of sequences similar to the known TFBS differed in the set of coexpressed genes compared to all genes in the genome. The hypothesis was that genes in the coexpressed set were more likely to have a sequence similar to the TFBS in their upstream regions than genes in general.

Here, we describe the bioinformatic methods used in these two studies in detail. We expect that the comparison of these two studies will provide new insight into the investigation of gene regulation in coexpressed gene sets. First we will review the anatomy and development of the cell types in question, then examine the SAGE data. The bioinformatic results for each investigation will be summarized separately, and the conclusions will be discussed jointly.

2.1.1 Anatomy and Cell Development

2.1.1.1 The *C. elegans* Intestine

The *C. elegans* intestine is clonally derived from the E cell at the 8-cell stage of the *C. elegans* developing embryo [3,4]. In the adult worm, the intestine consists of 20 large epithelial cells [5] which comprise one-third of the worm's somatic cell body mass (Figure 2.1) [6].

[1] A version of this chapter will be submitted for publication. Monica C. Sleumer, Mikhail Bilenky, Gordon Robertson, John F. Etchberger, Adam Lorch, James D. McGhee, Oliver Hobert, Donald G. Moerman, Steven J. Jones. **Bioinformatic Methods for Investigation of Gene Regulation in Coexpressed Gene Sets.**

The intestine is the powerhouse of the worm and drives all of its other abilities, especially locomotion and reproduction. The anterior portion is primarily responsible for enzymatic digestion of ingested food, while the posterior portion is responsible for nutrient absorption and storage [5]. There are a wide variety of proteins expressed in intestinal cells. Central to the digestive function of the intestine are lysozymes to destroy the bacterial cell wall, ATPases to break down the macronutrients released from the ground bacteria that enter the intestine from the pharynx [6], and enzymes such as proteases, nucleases, and phosphatases. The intestine also expresses the genes for transport and metabolic proteins such as glycosyltransferases and lipases involved in fatty acid metabolism [5]. Because intestinal cells require a large number of different proteins to function properly, they also express housekeeping genes required for translation, RNA processing, transcription, and chromatin organization at high levels [6].

2.1.1.2 The *C. elegans* Nervous System and ASE Gustatory Neurons

C. elegans has a relatively complex nervous system by which it senses and responds to the environment; it contains 302 neurons in total in the *C. elegans* hermaphrodite [7]. The nervous system consists of three types of neurons: motor neurons, interneurons, and sensory neurons, and their connectivity and function are invariant between worm individuals. Motor neurons activate rhythmic muscular contraction in body wall muscles, intestinal muscles, and vulval muscles, making it possible for the worm to move, defecate, and lay eggs. Interneurons link sensory input to action and are found in the central nervous system: the nerve ring in the head and the ventral nerve cord [8]. Sensory neurons can be separated into two types: ciliated and nonciliated. Nonciliated sensory neurons include the gentle touch mechanosensory neurons, while all thermosensory (temperature), olfactory (smell), and gustatory (taste) neurons are ciliated [9].

Ciliated sensory neurons typically occur in pairs, and are stimulated by their respective external cues via the tips of their cilia. Many cilia are exposed to the environment via the amphids, two pores on either side of the mouth [10]. Other ciliated sensory neurons reach the environment through the lips of the mouth and through phasmids (also pores) in the tail. The cilia of the thermosensory neurons are not exposed directly to the outside environment [11].

ASE neurons are gustatory neurons that allow the worm to detect water-soluble edible chemicals such as salts, amino acids, and small metabolites (Figure 2.2). There are two neurons in this class, ASER and ASEL, utilizing the right and left amphids respectively. The two neurons have a slight difference in chemoreceptor composition; the primary difference between them is that ASER expresses the gene *gcy-5* while ASEL expresses *gcy-7* and *lim-6* [12]. When ASE neurons are destroyed by laser ablation, the worm loses all ability to sense and move towards food in its environment, a process called chemotaxis [13]. The zinc finger TF CHE-1 is necessary for the expression of ASE-specific markers; a *che-1* deletion mutant has a phenotype similar to that of a worm whose ASE neurons were destroyed [14].

2.1.2 Tissue-specific Expression

Serial Analysis of Gene Expression (SAGE) is a laboratory technique used to measure gene expression in a specific tissue [15]. In this method, short sequences (called "tags") are extracted from each polyadenylated RNA, concatenated, and sequenced. The tags are used to identify the genes that are expressed in that tissue, and the number of occurrences of each tag (the "tag count") indicates the expression level.

SAGE has been used to make a large number of libraries for a wide variety of *C. elegans* tissues under different conditions [16]. In both studies discussed here, we compared two SAGE libraries which had only one difference in the conditions in which they were made, and identified genes that were expressed much more strongly in one library than the other. All four SAGE libraries were created by the British Columbia *C. elegans* Gene Expression Consortium.

2.1.2.1 Tissue-specific Expression in the Intestine

For the Intestinal study, we compared gene expression, as measured by SAGE, in hand-dissected *C. elegans* intestinal tissue (the intestinal library) to gene expression in whole worm (the somatic library). The somatic *C. elegans* tissue was estimated to consist of 1/3 intestine, and both SAGE libraries were approximately the same size in terms of total tag count, so an intestinal:somatic tag count ratio of three or more meant that the gene was expressed primarily in the intestine.

We identified 74 genes that were specifically expressed in the *C. elegans* intestine (Table 2.1). These genes had a very high expression in the Intestinal library (tag count greater than 50) and also had an Intestinal:Somatic tag count ratio greater than three. Additionally, ribosomal genes and genes with short upstream regions (suspected to be in the downstream position in an operon) were excluded from the list.

We hypothesized that the transcription factor ELT-2 was the primary TF responsible for embryonic development and maintenance of gene expression in the adult worm intestine. ELT-2 is known to be a GATA-type zinc finger factor via orthology to other GATA factors. Although the *C. elegans* genome contains two other intestinally-expressed GATA factors, ELT-4 and ELT-

7, *elt-2* is an essential gene while *elt-4* and *elt-7* are not: *elt-2* loss-of-function mutants have malformed intestines and die shortly after hatching [17].

The gene *elt-2* is first expressed during embryonic development of the intestine [5]. We assembled a list of experimentally validated ELT-2 binding sites that were shown to be necessary for intestinal expression of their respective genes (Table 2.2). We did not find any examples of intestinally-expressed genes that were regulated by any other transcription factor. In order to see the sequence conservation in each position, we created a sequence conservation logo [18] from the list of ELT-2 binding sites that showed they were characterized by a central TGATAR motif (Figure 2.3).

2.1.2.2 Tissue-specific Expression in the ASER Neuron

For the ASE Neuron study, we compared gene expression in the ASER neuron to gene expression in the AFD neuron, a thermosensory neuron. The ASER SAGE library had been generated by taking advantage of the fact that the ASER neuron is the only cell in *C. elegans* to express the gene *gcy-5* [12]. Worm embryos were injected with a DNA construct containing the *gcy-5* promoter coupled to the green fluorescent protein (GFP) gene. As the embryos developed, only the ASER neurons expressed GFP, even though all cells in the embryo contained the DNA construct. The embryos were then disrupted and their cells sorted using a fluorescent-activated cell sorter (FACS). The GFP-expressing cells, which were all immature ASER neurons, were then processed using the SAGE technique in order to obtain a gene expression profile. A second SAGE library was made from pure AFD thermosensory neurons using a similar strategy; AFD neurons are the only cells that express the gene *gcy-8* [19], and therefore a DNA construct consisting of the *gcy-8* promoter coupled to GFP caused only the AFD neurons to fluoresce.

We identified several categories of genes that were expressed more strongly in the ASER neuron compared to the AFD neurons. These included TFs, sensory receptors, and signaling proteins such as neurotransmitters and adhesion molecules. In total, we identified 1302 genes that appeared to be important to ASER cells (referred to as the ASE>AFD gene set), and we retested the expression of 49 of these by creating promoter-GFP DNA constructs, injecting them into worm embryos, and observing the subsequent expression of GFP. Seventy percent of them showed strong expression in ASE neurons.

We identified CHE-1 to be the TF responsible for regulating gene expression in ASE neurons and characterized the CHE-1 binding site. For several of the genes which we had previously identified as being expressed in ASE neurons, we performed a series of deletion mutagenesis and GFP-construct experiments to narrow down the exact portion of each gene's promoter that was necessary to maintain ASE expression. Some of these genes were previously known to be regulated by CHE-1, and we showed that all of the GFP expression was lost in *che-1* loss-of-function mutant worms. We then performed competitive electrophoretic mobility shift assays to confirm that CHE-1 bound the sequences we had identified as necessary for ASE expression. Finally, we aligned all of the CHE-1 binding sites and observed that they had similar sequences (Table 2.3). We observed that these sites were similar to both the predicted CHE-1 binding site [20] and to experimentally validated sites of the *Drosophila* CHE-1 orthologue GLASS [21,22].

Just as for the ELT-2 binding sequences, we generated a sequence conservation logo [18] of the CHE-1 binding site (Figure 2.4). The site had four invariant base positions and two highly conserved positions.

2.1.3 Objectives and Approach

For the Intestinal study, we knew that ELT-2 was responsible for the regulation of many intestinal genes, but we had few examples of ELT-2 binding sites. Our bioinformatics objectives were to determine which motifs were prominent in the upstream regions of the 74 transcripts (and their orthologues) we had identified as intestine-specific, form one or more position frequency matrices (PFMs) based on the motif discovery results, and scan the upstream regions of both intestinal and non-intestinal genes with the PFMs and determine whether there was a difference in the score distribution of the motif(s) in the various sets.

For the ASE Neuron study, our bioinformatics objective was to find out whether genes that were specifically expressed in the ASE neuron were more likely to harbour a high-scoring CHE-1-like site, and whether the level of expression in ASER was related to the proportion of genes that had a high-scoring CHE-1-like site.

A similar approach was used to address both sets of questions (Figure 2.5). In both cases, we compared two SAGE libraries to obtain sets of genes that were more strongly expressed in one tissue than the other tissue. We then used experimentally validated TFBSs to form a PFM and scanned the upstream regions of *C. elegans* genes for sequences that matched the PFM. Finally, we graphed the cumulative distribution of the maximum-scoring match to the PFM from each upstream region and showed that genes that were more strongly expressed in the tissue of interest were more likely to have a high-scoring PFM match in their upstream regions (Figure 2.5, paired black and red arrows). For the Intestinal study, we also used a motif discovery step to obtain more examples of sequences resembling the TFBS and used them to augment the PFM (Figure 2.5, solitary black arrows). For the ASE Neuron study, we performed a statistical analysis on the cumulative distributions to determine statistical significance (Figure 2.5, solitary red arrow).

For the Intestinal study, we used two motif discovery algorithms, MotifSampler [23] and RSAT [24], and combined the results. It has been shown that some regulatory elements are conserved among species in the upstream regions of orthologous genes [25,26]. Therefore, we obtained orthologues for the 74 specifically-expressed genes in *C. briggsae* and *C. remanei* using the WABA alignment algorithm [27] and repeated the motif discovery procedure for the upstream regions of the orthologues. We also applied the same motif discovery methods to three sets of 74 randomly chosen genes from the *C. elegans* genome to determine the significance of motifs found in the specifically-expressed set.

Motifs obtained from the two motif discovery programs were combined and then clustered using the OPTICS clustering algorithm [28]. A PFM was made from the sequences in the largest cluster by counting the frequency of occurrence of each base in each position of the motif and normalizing to make the sum of each column equal to one. We then used the PFM to score sequences by simply summing the values in the matrix for the bases corresponding to the sequence in question.

From the PFM we created a sequence search pattern representing all nucleotide variations seen at each position of the motif. Some positions of the motif were invariant – only one nucleotide was seen – while in other positions only some nucleotides were seen. We scanned the upstream regions of all protein-coding transcripts in the *C. elegans* genome for matches to the pattern. Each sequence in each upstream region that matched the pattern was scored and recorded; we retained only the highest-scoring match in each upstream region. We then graphed the cumulative distribution function of the highest-scoring match in each upstream region in *C. elegans*, and compared this distribution to that of various subsets of genes (e.g. genes that were strongly expressed in the intestine or ASE neuron) obtained from the SAGE data. For the ASE Neuron study, we used the Kolmogorov-Smirnov test to determine whether the distributions of the gene subsets were significantly different from that of all genes. For those distributions that were different, we had shown that the level of gene expression in the tissue of interest was related to the likelihood of having a high-scoring motif in the upstream region.

2.2 Results – Intestinal Study

2.2.1 Motif Discovery and Clustering

We used MotifSampler and RSAT to find motifs of width 6, 8, 10, and 12 bp in the upstream regions of the 74 intestinal genes and obtained 204 motif instances (Figure 2.6). The motif discovery programs report motifs that occur more often in the foreground set than in the background. However, they do not report which motifs are similar to each other and how many different types of motifs there are. Therefore we applied a sequence alignment program followed by a clustering algorithm to the discovered motifs to place them into clusters of similar sequences.

ClustalW [29] was used to align motifs with respect to each other. Flanking sequence was added to each motif to make all aligned sequences the same width, resulting in a column of sequences 23 bases wide. A simple mismatch counter was used as a distance function between aligned motifs, and the OPTICS hierarchical clustering algorithm [28] was used to place the resulting motifs into clusters. OPTICS indicated that there were two main signals in the motif discovery result, a large cluster consisting of 111 sequences and a small cluster consisting of 6 sequences. The largest cluster featured a conserved TGATAA sequence (Figure 2.7), which strongly resembled the ELT-2 binding sites we had previously collected (Figure 2.3).

Orthologues for the 74 *C. elegans* genes were identified in the genomes of *C. briggsae* and *C. remanei* using the WABA alignment algorithm [27], and the motif discovery, alignment, and clustering procedures were repeated for the upstream regions of these orthologues. For *C. briggsae*, orthologues were found for 57 of the genes while 39 orthologues were found in the *C. remanei* genome. In both cases, two clusters of motifs were formed, a large cluster resembling an ELT-2 binding site, and a small cluster that did not resemble anything seen in the other species (Figure 2.8).

In the three sets of randomly chosen negative controls, no clusters of motifs of size greater than five were found. This showed that the small clusters of motifs were not significant; any set of genes could produce a weak, sparse motif by chance. Therefore, small clusters of motifs in all three species were ignored.

2.2.2 Motif Distribution

We combined the sequences in the large cluster of motifs discovered in the 74 *C. elegans* upstream regions (Figure 2.7) with the experimentally validated TFBSs (Table 2.2 and Figure 2.3), a total of 127 sequences. The base frequencies of the central 10 most-conserved bases were tabulated and normalized to form a position frequency matrix (PFM) (Table 2.4).

In order to visualize the level of conservation in each column of the PFM, we created a sequence conservation logo (Figure 2.9). All of these sequences were variations on the pattern NNNGATARNN except one. Therefore, we made the assumption that other ELT-2 binding sites would also match this pattern. We scanned the upstream regions of all protein-coding transcripts in the *C. elegans* genome and scored each pattern match by summing the respective frequencies from the PFM and normalizing by the width of the pattern (10 bases), retaining only the highest-scoring match from each upstream region.

The lowest score that could be obtained from a matching sequence was 0.44, while the highest possible score was 0.8. The highest-scoring experimentally validated site had the maximum score of 0.8 (*cpr-1*), and the lowest-scoring experimentally validated site had a score of 0.53 (*smd-1*), but it was one of three ELT-2 sites upstream of that gene (see Table 2.2). The lowest-scoring freestanding ELT-2 binding site had a score of 0.64 (*spl-1*).

We graphed the cumulative distribution of the highest-scoring match of all protein-coding transcripts in the genome (Figure 2.10, black line). The results showed that 70% of *C. elegans* transcripts had a pattern match with a score of 0.65 or higher. Looking at only the 2816 genes with a tag count greater than one in the intestinal SAGE library, we saw that the distribution of pattern matches was shifted to the right (Figure 2.10, magenta), indicating that these genes were more likely to have a high-scoring pattern match in their promoters. The 291 genes that had a high expression in the intestinal SAGE library (count greater than eight) - and were also detected in the somatic cell library at roughly the same expression level - had a similar distribution (Figure 2.10, green). Genes that had a much higher expression level in the intestinal SAGE library than the somatic cell library had a distribution shifted even further to the right (Figure 2.10, blue). Almost 85% of these 534 genes had a pattern match with a score of 0.65 or higher.

By contrast, a set of 33 ribosomal protein-coding genes, which were highly expressed in the intestine but did not have intestine-specific expression, had a distribution that was shifted to the left (Figure 2.10, orange), indicating that these genes were less likely to have a high-scoring pattern match in their upstream regions. Lastly, all 74 of the intestine-specific genes had a pattern match in their upstream regions, 90% of them had a pattern match with a score > 0.7, and almost 30% of them had a pattern match with a score of > 0.79 (Figure 2.10, red). As a negative control, we randomly selected and graphed the cumulative distribution function (CDF) of 100 sets of 74 genes (Figure 2.10, cyan). The resulting curves formed a wide band around the whole genome curve, but we observed that the intestine-specific curve was far outside of this band, showing that the distribution is highly significant and could not have occurred by chance.

2.3 Results – ASE Neuron Study

For the ASE Neuron study, we assembled a set of experimentally validated CHE-1 binding sites (Table 2.3) and made a PFM from them using the same normalization procedure we used for the Intestinal study (Table 2.5). Our objective was to analyze the distribution of sequences matching a pattern formed from the CHE-1 sites to see whether they were associated with gene expression in the ASE neuron.

The minimal pattern that matched all of the experimentally validated sites was GAADCMNHNNNH, and we used the same scanning technique with this pattern that we had used for the ELT-2 site pattern scan. The minimum possible score for a pattern-matching sequence was 0.4 and the maximum score was 0.73. Among the experimentally validated sites in Table 2.3, the highest-scoring site had a score of 0.72 (*ceh-36*), and the lowest-scoring site had a score of 0.57 (*lim-6*), but it is one of two CHE-1 sites upstream of that gene. The lowest-scoring freestanding site had a score of 0.61 (*nlp-3*).

Using the same method as for the Intestinal study, we graphed the cumulative distribution of the highest-scoring match of all protein-coding transcripts in the genome (Figure 2.11, black curve). We observed that it formed a sinusoidal curve; 20% of transcripts had no match at all, and 30% of transcripts had a match with a score > 0.6. The first subset of genes that we looked at in comparison to this curve was the ASE>AFD gene set that we had initially identified as being important to ASER neurons. The scan produced results for 1273 of the genes (the others were not protein-coding), and the resulting curve was shifted only slightly down from the whole-genome curve (Figure 2.11, purple).

The curve for genes with a SAGE tag count > 4 in the ASER SAGE library and 0 in the AFD library was shifted to the right near the top of the curve, indicating that these genes were more likely to have a high-scoring match (Figure 2.11, magenta), while the curve for genes with a tag count > 4 in the AFD library and not observed in the ASE library was not shifted (Figure 2.11, blue). Similarly, the curve for genes with a tag count > 6 in the ASE library and 0 in the AFD library was shifted even further to the right (Figure 2.11, grey).

The curve for genes that were observed in both libraries but had an ASE:AFD tag count ratio > 5 was shifted downwards near the bottom of the curve (Figure 2.11, red), while the curve for genes with an AFD:ASE tag count ratio ≥ 5 was not shifted (Figure 2.11, green). Similarly, the curve for genes with an ASE:AFD tag count ratio ≥ 7 was shifted even further down (Figure 2.11, orange).

The CDF of the scores from the 27 experimentally validated CHE-1 sites from Table 2.3 is shown on the graph in yellow (Figure 2.11). All of these have a score of 0.57 or higher, so the resulting curve is not near the curves from the other gene sets. In order to confirm that the scan was able to reproduce the experimental results, we also plotted the scan results from the genes with experimentally validated sites (Figure 2.11, brown). There were only 24 results from this scan because we only used the maximum-scoring match from each of the genes, and one of the 25 genes (*lsy-6*) was a microRNA gene rather than a protein-coding gene. For several genes, the scan found a different top-scoring motif than the experimentally validated site, either because the experimentally validated site was so far upstream that it was out of range, or because the scan found a match that had an even higher score than the experimentally validated site. However, for one of the genes (*gcy-6*), no match was found; the experimentally validated site was beyond the next-most upstream gene (*trx-2*). Finally, we randomly sampled 100 sets of 100 genes from the whole genome set to serve as negative controls. The curves for these sets formed a narrow band around the whole genome curve (Figure 2.11, cyan).

We used the Kolmogorov-Smirnov test to determine which of the curves had a statistically significant difference in distribution compared to the whole genome curve. The difference in the distribution of the ASE>AFD gene set was not quite statistically significant (p=0.0597). Both of the curves from genes strongly and exclusively expressed in the ASE neuron (ASE tag count > 4 and ASE tag count > 6), and both curves from genes strongly overrepresented in the ASE neuron compared to the AFD neuron (tag count ratio ≥ 5 and ≥ 7) had significant p-values (p= 0.0486, 0.0309, 0.009, and 0.0089 respectively). By contrast, genes strongly expressed or overrepresented in the AFD neuron in comparison to the ASE neuron did not have statistically significant differences in their score distribution. Of the 100 negative control sets, two had p-values between 0.015 and 0.05, which is what we would expect to see by chance.

2.4 Discussion and Conclusions

2.4.1 Intestinal Study Discussion

MotifSampler uses a Gibbs sampling algorithm that finds overrepresented motifs in a stochastic process. The advantage of using MotifSampler is that it finds whole motifs, including common sequences and minor variations on common sequences. However, the finding of each instance of a motif is dependent on the genomic context. If a common sequence is in the middle of a region of low complexity, MotifSampler will ignore it, even if it is identical to other instances. By contrast, RSAT uses a word-counting algorithm to obtain a complete count of all n-mers in both the foreground and background sequences and finds overrepresented n-mers rather than motifs. The advantage of RSAT is that it finds all instances of overrepresented sequences regardless of their genomic context, but the disadvantage is that each sequence is assessed individually; sequence variations are not included unless they are found independently.

Another limitation of RSAT is that it has a maximum motif width of 8 bp, with the result that it may miss wider overrepresented sites. Combining the results from the two algorithms mitigated the disadvantages of each method and ensured that almost all interesting sequences were located (Figure 2.6).

The TGATAR motif cluster signal was weaker in *C. briggsae* than in *C. elegans*, and weaker still in *C. remanei* (Figures 2.7 and 2.8). Several factors may have contributed to the signal dilution. For example, not all of the 74 *C. elegans* upstreams contained a detectable TGATAR signal to begin with (70 of them had an instance of a motif from the motif discovery; 59 of them had a signal strong enough to be used in the PFM). Not all of the genes had clearly identifiable orthologues, so some of the ELT-2-like sites in the other two species were not detected; for others, the orthologues may have been assigned incorrectly, in spite of our rigorous requirements, resulting in the incorporation of an upstream region that was not intestinally expressed. Additionally, we may not have identified the translation start sites of the orthologous genes accurately, so some of the orthologous ELT-2 sites may have been out of range of our search. Lastly, it is possible that *C. briggsae* and *C. remanei* have evolved to regulate some of the genes using something other than the ELT-2 TF. It is unlikely that the ELT-2 orthologue in *C. briggsae* and *C. remanei* binds to a different sequence in these two species, because GATA factors are conserved from yeast to mammals.

We demonstrated the existence of several sites with a score of 0.66 that do not bind ELT-2 and do not have an impact on gene regulation, and we hypothesized that the biologically relevant pattern matches to the PFM (that is, validated ELT-2 binding sites) are likely to lie in the range from just above 0.65 to 0.80. We observed that ELT-2-like sites were very common in the genome – about 70% of genes had a site above this threshold in their upstream regions (Figure 2.10). Some of these sites may be functional binding sites for one of the other nine GATA TFs in the *C. elegans* genome such as END-1 and END-3, which are known to regulate genes expression in the hypodermis [30]. However, ELT-2-like sites will occur by chance in the genome, and the pattern-match scan will find all sequences similar to an ELT-2 site whether they are functional or not, so most of the sites are not functional TFBSs. It is known that scanning for sites using a PFM will yield > 99% false positives [31], but our purpose here was not to find functional ELT-2 binding sites, only to demonstrate the relative distribution of ELT-2-like sites among intestine-expressed, intestine-enriched, and intestine-specific genes as compared to all genes.

Conversely, about 15% of intestine-enriched genes did not have an ELT-2-like site in their immediate upstream regions (Figure 2.10). Some intestinal genes may be regulated by one or more of the 108 intestine-enriched TFs that we identified instead of by ELT-2 directly. It makes sense that a worm would be able to fine-tune the expression of some intestinal genes depending on its diet and environmental conditions. We also identified a number of genes that are regulated by ELT-2 in conjunction with another TF. For example, ELT-2 and SKN-1 may jointly regulate antioxidant genes under stress conditions [32], and the metal-response gene *mtl-2* is activated by ELT-2 but repressed by an unknown TF in the absence of toxic metals [33].

The fact that other TFs must be involved in the regulation of at least some intestinal genes leads us to the question of why the motif discovery procedure only found ELT-2 like sites. We only used genes with strong, intestine-specific expression for the motif discovery step. We reported the single strong signal that we found and did not continue to search for weaker motifs. We briefly considered searching the upstream regions for nuclear hormone receptor (NHR) binding sites because 15 of the TFs expressed at high levels in the intestine were NHRs. However, NHRs have highly variable binding sites that can not be detected easily using motif discovery [34]. In order to detect the binding sites of other secondary TFs that control intestinal expression, we would need more specific gene expression data, such as a subset of intestinally-expressed genes that are all upregulated under a specific dietary or environmental condition.

2.4.2 Intestinal Study Conclusions

We have shown that the set of intestine-specific genes was sufficient for the computational identification of a functional TFBS: the motif discovery results were concordant with the previously identified binding sites. We have also shown that the PFM was a valid model of the ELT-2 binding site: the likelihood of having a high-scoring ELT-2 binding site in the upstream region was related to the level and specificity of expression in the intestine. Both widely-expressed and intestine-enriched genes were more likely to have high-scoring ELT-2-like sites in their upstream regions, suggesting that ELT-2 is responsible for the intestinal expression of genes that are also expressed in other tissues.

ELT-2 is the primary TF responsible for expression in intestinal genes, except housekeeping genes such as ribosomal proteins. The ELT-2 binding site was the only significant signal produced by the motif discovery procedure, and most intestinally-expressed genes have a high-scoring ELT-2-like site in their upstream regions. Other TFs may be used in conjunction with - or secondarily to - ELT-2 under specific environmental conditions.

Gene regulation in the intestine is evolutionarily conserved in other species in the genus *Caenorhabditis*. The ELT-2-like site was the only significant signal produced by motif discovery in the upstream regions of *C. briggsae* and *C. remanei* orthologues of the intestine-specific genes, providing further evidence for the predominance of ELT-2 as a regulator of gene expression in the nematode intestine.

2.4.3 ASE Neuron Study Discussion

2.4.3.1 Similarities and Differences between ASER and ASEL

The primary SAGE library used in the ASE Neuron study was made from embryonic FACS-sorted cells expressing GFP from the *gcy-5* promoter. During embryogenesis, cells divide and differentiate in an invariant process, and the ASE neuron on the right side of the embryonic worm always takes on the characteristics of the ASER neuron while the ASE neuron on the left becomes the ASEL neuron [4]. Both ASE neurons are gustatory, but once they mature, they have different sensitivities to salt ions [35] and express different chemoreceptors and neurotransmitters [7]. ASE neurons are born 350 min after fertilization, and remain in a hybrid state until 500 min after fertilization during which they both express some of the fate markers that are later expressed exclusively by either ASER or ASEL [12]. However, *gcy-5* is exclusively expressed in the ASER cell at all stages of development [12], therefore, the SAGE library was made from > 90% pure ASER neurons, not from a mixture of both ASE neurons. GFP construct experiments performed on genes from the ASER SAGE library showed expression in both neurons, which implies that most genes in the ASE>AFD list are expressed in both ASE neurons.

ASER and ASEL have completely separate lineages going back to the 4-cell stage (ASEL derives from the ABa cell and ASER derives from the ABp cell); they do not both result from a single ASE precursor (Figure 2.12) [4]. Therefore, they must converge onto similar expression and function immediately after they are born, and then go on to diverge into ASER and ASEL as the embryo matures. Because they both synapse onto some of the same interneurons [36], it may be that the different neurotransmitters made by the two ASE neurons are used by the postsynaptic interneurons to distinguish which neuron the signal is coming from.

The gene *che-1* was already known to affect both ASER and ASEL – knocking out *che-1* destroys the function of both neurons [14]. We showed that promoters of both ASER-specific (*gcy-5*), and ASEL-specific (*gcy-7* and *lim-6*) genes contain CHE-1 binding sites and require CHE-1 for expression. We hypothesized that both ASE neurons switch *che-1* on shortly after they are born, and begin to express ASE genes, including some that later become exclusive to ASER or ASEL. They then complete the differentiation process via a bistable feedback loop in which expression of inappropriate genes for each neuron are repressed. The origin of the switch in the bistable feedback loop is unknown, but it may already be present in the cell before differentiation occurs, because the lineage of ASER involves various other neurons on the right side of the worm and the lineage of ASEL involves the corresponding left-sided neurons (Figure 2.12) [4].

This could be investigated by examining the gene expression of other neurons related to ASE by lineage. First we would need to determine left/right specific genes for each neuron pair, analogous to *gcy-5* and *gcy-7* for the ASE neurons. We could then generate GFP expression constructs and SAGE libraries from individual FACS-sorted cells just as was already done for the ASER neurons. From the SAGE data we would be able to find factors that left-sided neurons have in common with each other but not with the right-sided neurons; we could also investigate the upstream regions of the pooled left-specific genes and see if they have any motifs in common, then repeat the investigation with the right-specific genes. The AFD SAGE library would not be suitable for this type of analysis because it was made from both AFDR and AFDL.

2.4.3.2 Summary of Functional Evidence of CHE-1 Binding Sites

The approximate binding site of the CHE-1 TF was first predicted using the zinc finger binding site prediction generator C2H2-enoLOGOS [20]. The resulting prediction was strikingly similar to the eventual experimentally validated site and provided encouraging evidence for the function of the site, but was not specific enough to do *ab initio* scanning. Additionally, we found that two binding sites of the *Drosophila* TF GLASS [21,22], which has 100% amino acid identity with CHE-1 in the zinc fingers of the DNA binding domain, were very similar to those that we found for CHE-1.

We found that the CHE-1 sites usually occurred in a single copy, primarily within 1 kb of the gene's ATG, but occasionally further upstream and even beyond the next-most upstream gene. The binding sites functioned in an orientation-independent manner and were found on both strands of the DNA sequence (relative to the strand of the gene being regulated). We showed that the CHE-1 binding sites were necessary for ASE expression: deletion of the sites also prevented expression in ASE for all 17 genes tested. Similarly, the CHE-1 binding sites were found to be sufficient for ASE expression: we were able to restore ASE expression from a severely truncated promoter of *ceh-36* by adding only a CHE-1 binding site. We were also able to induce ASE expression from a promoter that normally has no transcriptional activity at all by inserting eight concatenated CHE-1 sites.

Lastly, in the long series of deletion mutagenesis experiments aimed at determining regulatory elements for ASE-expressed genes, we only found one instance of a site that did not match the CHE-1 site. The non-matching site was upstream of *lim-6*, a homeobox-domain TF, and we hypothesize that it may be an autoregulatory element responsible for ASEL-specific expression of *lim-6*.

2.4.3.3 Upstream Region Pattern Scan

Comparison of experimental sites to the pattern-match scan of the upstream regions of the same genes provided a way to assess the accuracy of the scan results with respect to physiological CHE-1 binding sites (Figure 2.11). The similarity of the brown and yellow curves showed that results of the pattern scan provided a reasonable facsimile of the deletion mutagenesis results. However, in several cases, the scan found a different CHE-1-like site than the one that was found using laboratory methods. In three cases, the scan found a site that had an even higher score than experimentally validated site; the function of these sites is unknown. In three other cases, the experimentally validated site was out of range of the scan; for two of these, the scan found a lower-scoring pattern match (biological function unknown once again), and for the third upstream region no other match was found. In this last case, the gene *gcy-6*, the CHE-1 binding site was in fact so far upstream that there was an entire gene (on the opposite strand) between the TFBS and the regulatory target. It is important to keep in mind that although the immediate upstream regions of genes are clearly enriched for regulatory elements compared to the rest of the genome, in practice such elements can be found in a variety of other locations.

Although the CHE-1 TF was clearly essential to the function of ASE neurons, not all ASE-expressed genes had CHE-1 sites in their immediate upstream regions. For some of them, the CHE-1 site was simply out of range of the scan; others may have been regulated by one or more of the 68 TFs that were expressed at a higher level in the ASE neurons than in the AFD neurons. These TFs could be used to fine-tune gene expression in the ASE neuron under different environmental conditions. For example, *flp-4* has been shown to be expressed in the ASE neurons [12], and we found that its expression was CHE-1 dependent, but its upstream region contained no CHE-1 binding site. Other genes in the ASE>AFD list may not have been ASE-specific and may have been regulated by factors common to neurons including ASE but excluding AFD.

Conversely, some genes not expressed in ASE seemed to have high-scoring CHE-1-like sites in their upstream regions. One of these, upstream of the serpentine receptor *srt-63*, was tested and found to function as a CHE-1 site on its own and in the context of the *gcy-5* promoter but not in the context of the *srt-63* promoter. Promoters with CHE-1-like sites that are not expressed in ASE may contain binding sites for other TFs that prevent CHE-1 binding. Further scanning deletion mutagenesis experiments will need to be performed on the regulatory regions of genes with inactive ASE motifs to clarify what other sequences and factors determine the functionality of the ASE motif.

Interestingly, for sets of genes with very high expression in ASE cells and no expression in AFD (ASE only, > 4 and > 6 tag counts, magenta and grey), maximum separation from the whole genome curve occurred around a score of 0.62, indicating that these sets had a much higher proportion of high-scoring motifs than all genes in the genome (with p-values of 0.0486 and 0.0309 respectively). By contrast, for sets of genes with very high expression in ASE cells relative to AFD cells (ASE>AFD 5x and 7x, red and orange), maximum separation occurred at a score of around 0.5, indicating that these sets had a much lower proportion of genes with no motif at all (with p-values of 0.009 and 0.0089 respectively).

At least one example was found of a microRNA gene that was regulated by CHE-1 (*lsy-6*). This gene together with another microRNA gene (*mir-273*) has previously been shown to be central to the bistable switch that governs the difference in gene expression between ASER and ASEL [12]. However, our pattern-match scan did not find the CHE-1 site upstream of *lsy-6* because we limited our analysis to protein-coding genes on the basis that they are better understood and their genomic boundaries better defined. The importance of microRNAs to the gene regulation of ASE neurons shows that microRNAs should be taken into consideration in future analyses, both as targets of regulation by TFs and as mechanisms of gene regulation.

2.4.4 ASE Neuron Study Conclusions

We have shown that the comparison of two SAGE libraries made from FACS-sorted cells is a powerful technique to identify tissue-specific genes. We identified transcripts specific to ASE gustatory neurons (as compared to AFD thermosensory neurons), including protein-coding transcripts, microRNAs, and antisense RNAs, and showed that a wide variety of chemoreceptors, TFs, and neurotransmitters are expressed in ASE neurons.

CHE-1 is the primary, but not the only, TF that regulates ASE-specific expression. ASE-expressed genes were significantly more likely to have a high-scoring CHE-1 like site in their upstream regions; the higher the genes' expression in ASE, the more significant the difference was. About half of the ASE-expressed genes did not have a CHE-1 binding site in their immediate upstream regions. Some of these may be directly regulated by CHE-1 but have a CHE-1 binding site outside of the search zone. Others may be regulated by one of the other 78

TFs expressed in the ASE neurons, including TFs whose expression is activated by CHE-1 and TF(s) that were responsible for the initial activation of CHE-1 expression.

The CHE-1 binding site is both necessary and sufficient for expression in ASE neurons for many genes. However, the function of CHE-1 binding sites is context-dependent, and this dependence is not fully understood: many genes have a CHE-1-like site in their upstream regions but are not expressed in ASE.

2.4.5 Overall Discussion

In both the Intestinal study and the ASE neuron study, we formed a TF binding model from a list of putative TFBSs. In the Intestinal study, the list originated from the results of a motif discovery procedure, while for the ASE neuron study, it originated from a series of experimentally validated TFBSs. The advantages of using motif discovery are that it is less time- and resource-consuming than laboratory research and can be used to leverage a large quantity of gene expression data to find novel TFBSs. However, in order for motif discovery to be successful, gene expression data, generated under specific conditions, is required to produce a reasonably-sized list of co-regulated genes. Additionally, TFs responsible for regulation of genes must have specific binding sites whose similarities are computationally detectable. Highly degenerate TFBSs, binding sites with no consistent width, and sites that appear frequently in the genome can not be found using computational methods. Even when all requirements are met, as they were in the Intestinal study, motif discovery can not distinguish between functional binding sites and DNA sequences that are similar to the TFBS but are not bound by the TF. It is likely that some of the examples we used for the ELT-2 binding model were not functional ELT-2 binding sites. In contrast, the advantage of using only experimentally validated binding sites is that the model will be more accurate and undiluted by nonfunctional sites. However, the time-consuming nature of deletion mutagenesis experiments means that there will be fewer examples of experimentally validated binding sites, and if there are too few examples, the binding model will not be able to successfully distinguish significant score distributions. Therefore, the best binding model will be made from a combination of experimentally validated sites and motif discovery results.

For both studies, the statistically significant differences in the distributions of the gene sets showed that the scanning procedure was a useful way to assess the validity of the binding model and simultaneously explore the association between the TFBS and gene regulation in the tissue in question. However, the PFM binding model is not infallible and it is unlikely that all of the pattern matches, including the high-scoring ones, are functional binding sites. The comparison of the scan results with the experimentally validated CHE-1 sites for the same genes revealed that the scan picked up different sequences part of the time. Additionally, a sequence pattern scan (or any other bioinformatic technique) will necessarily have very strict parameters (length of upstream region, exclusion of coding sequence of nearby genes, maximum-scoring site only) that the actual physiological TF-DNA interaction does not have. We should keep the limitations of motif discovery and motif scanning in mind when we look at computationally-derived results and not assume that a high score (by whatever metric we are using) automatically translates to biological significance. Conversely, the lack of a significant computational result does not mean that the sequence in question has no function.

In both analyses we observed that many genes in the coexpression group did not have a binding motif or anything resembling it. There are three explanations for this finding. The first explanation is that the genes were not regulated by the primary TF we were looking at and there was no site to find; this possibility was already discussed for each study individually above. The second possibility is that there were matching sites, but they were outside the range of sequence we looked at. Other places where the sites might have been include: further upstream of the gene (including in the coding sequence, in the introns, or beyond an upstream gene), in the introns of the gene itself, or downstream of the gene. For this study we focussed our search on the immediate upstream region because it has been shown to be specifically enriched for TFBSs, but this necessarily meant that we missed some of them. The third explanation is that our PFM-based pattern scan did not accurately separate functional TFBSs from non-binding DNA sequence. The PFM was merely a model based on sequence similarity and did not take into account the energetics of TF-DNA binding and the mechanics of gene activation. It was entirely dependent on the input set of sequences; more and/or different input sequences would have changed the model and thereby changed the output from the scan. Some genes may have had several weak binding sites instead of one strong one, and some sequences sites may have strongly bound the TF in question even though they did not match the pattern or had a low score by our metric. We used a very simple binding model because it produced significant results, and because without more details of the biophysical properties of the TF-DNA binding relationship there was no clear way to improve on it. Overall, a limitation of the entire analysis is that we can not distinguish which of these three explanations was the correct one for each gene that did not have an associated pattern match.

In both analyses we found one major TFBS and showed that high-scoring sequences matching the TFBS were found in significantly higher proportions in the upstream regions of genes with enriched expression in the tissue of interest compared to the upstream regions of all genes. In both cases the TF we analyzed seemed to be the primary TF for that tissue; deletion of the gene for the TF caused complete failure of the development of the tissue. Both expression analyses showed that other TFs are also expressed in these tissues, and that these TFs may fine-tune the expression of some of the genes in the tissue. In order to find them we would need more specific data showing changes in gene expression in that tissue under different conditions.

The difference in distributions was greater for the ELT-2 analysis than for the CHE-1 analysis, and there are several possible reasons for this. First, ELT-2 had a more specific binding site: the average information content (a measure of conservation) of the ELT-2 PFM was 1.3 while the average information content of the CHE-1 binding site was 1.0. Secondly, we had more examples of experimentally validated ELT-2 and putative ELT-2 sites, so the resulting binding model was more accurate. Thirdly, the disparity between the ELT-2 and CHE-1 results may simply be an artifact of the possible scores between the pattern matches: there were 2048 possible 10bp sequences that matched the ELT-2 pattern and 13 824 12mers that matched the CHE-1 pattern. Lastly, it is possible that gene regulation in the intestine was (relatively speaking) broad and general with one primary controlling TF, while gene regulation in the ASE neuron was more complex, with fewer genes but more elements of transcriptional control.

2.4.6 Overall Conclusions

We have shown that SAGE libraries made from both FACS-sorted and microdissected *C. elegans* tissue are sufficiently accurate measures of gene expression to obtain sets of coexpressed and differentially expressed genes. A SAGE library made from a single tissue under normal environmental conditions provides a general picture of gene expression for that tissue.

Motif discovery in the upstream regions of coexpressed genes, or TFBSs found using deletion mutagenesis, or preferably a combination of both methods, can be used to form a valid TF binding model. This confirms that TFBSs are shared among coexpressed genes and implies that motifs identified in their upstream regions may be functional TFBSs, even if the binding protein is unknown.

In the absence of further information regarding the energetics of TF binding, a position frequency matrix is a valid model of a TFBS that is simple and makes no unwarranted assumptions. Scanning and scoring the upstream regions of genes for matches to the PFM can provide information regarding the distribution of sequences similar to the TFBS. Genes that were coexpressed in the tissue of interest were statistically more likely to have a match to the PFM compared to all genes in the genome. While the expression of some genes seemed to be controlled by a single TFBS, many other TFs are present in both intestine and ASE neurons, and regulation of other genes is clearly more complex. A more detailed understanding of the expression of those genes would require more specific gene expression data under a wide range of conditions.

In general, the results from both studies show that bioinformatics, high-throughput gene expression measurement methods, and laboratory research can be combined to obtain a multi-faceted understanding of gene regulation.

2.5 Methods

2.5.1 Motif Discovery

Upstream regions of the 74 genes in *C. elegans* (Table 2.1; genome sequence obtained from WormBase version WS140) and their orthologues in *C. briggsae* (from Cb25) and *C. remanei* (genomic sequence from the Genome Sciences Centre at Washington University in St. Louis) were collated into three species-specific files. Orthologues were determined using WABA [27]; only single WABA alignments that matched right from the ATG of the *C. elegans* sequence were retained. The upstream regions were taken as the lesser of 1500 bp (excluding masked repeats) or the distance to the end of the nearest upstream gene; a minimum upstream sequence length of 100bp was required. MotifSampler (widths 6, 8, 10 and 12 bp; [23]) and RSAT Oligo-analysis (width 8; [24]) were used to detect motifs. MotifSampler was run using the following parameters; -p 0.3 -s 1 -n 5 -r 100. The ‘r’ parameter specifies 100 iterations of Gibbs sampling; in order to reduce noise, only motifs that were detected in at least seven of the iterations were retained.

Species-specific backgrounds were generated for both methods. For MotifSampler, concatenated upstream regions from a large set of randomly chosen genes were used as the background, while for RSAT, counts of all 8mers in all upstream regions were used. Separate background counts were generated for each species using the RSAT custom background tool. All results were combined into one file. Overlapping motifs were merged into one, except in cases where there was a clear dimer: two motifs very close together.

Detected motifs were aligned with ClustalW [29] and clustered with OPTICS [28], using a base mismatch counter as a distance function between pairs of aligned motifs.

2.5.2 Scanning: ELT-2 Sites

Just as for the motif discovery procedure, upstream regions were defined as the lesser of 1500 bp upstream of the ATG (not including masked repeat sequence) or up until the end of the previous gene, whichever is shorter. A minimum upstream length of 100 bp was required to exclude genes in operons. DNA sequence was obtained from build WS140 of the *C. elegans* genome assembly, and the all-gene list was obtained from WormPep, using the 5'-most transcript in case of multiple transcripts for a gene.

A position frequency matrix (PFM) was generated using the combined results from the largest OPTICS cluster of motifs from *C. elegans* upstream regions and experimentally-determined sites (127 sequences in total; Table 2.4). All but one of these sequences were variations on the pattern NNNGATARNN (the exception was AATGATATAT). The upstream regions of all genes in the genome were scanned for instances of this pattern.

To attach a relative value to the quality of a match for a given sequence to the PFM, instances were scored by summing respective frequencies in each position and normalizing to the number of base pairs in the site (10). The maximum-scoring match for each upstream sequence region was recorded, and the cumulative distribution function of various sets of genes was graphed (Figure 2.10).

2.5.3 Scanning: CHE-1 Sites

We created a PFM from 27 ASE motifs that had been experimentally confirmed by EMSA to be CHE-1 binding sites (Table 2.5). All sites matched the pattern GAADMNHNNNH, and we scanned all upstream regions of protein-coding genes for matches to this pattern. The same set of *C. elegans* upstream regions was used, pattern matches were scored using the PFM, and the cumulative distribution functions were graphed using the same procedure as for the ELT-2 sites. Statistical significance of the difference between each set and the all-gene set was calculated using the Kolmogorov-Smirnov test as implemented by Matlab (Figure 2.11).

2.6 Chapter 2 Tables

B0218.8	*C03B1.12*	*C03G6.15*	*C05D2.8*	*C06B3.3*	*C07B5.5*	*C07D8.6*	*C08H9.2*
C10C5.4	*C14A6.1*	*C15C8.3*	*C28C12.5*	*C30G12.2*	*C34F11.3a*	*C39B10.3*	*C41C4.6*
C45G7.3	*C49C3.4*	*C50B6.7*	*C52E4.1*	*D2096.8*	*F14F4.3a*	*F15E11.12*	*F21F8.3*
F22A3.6a	*F28A12.4*	*F28D1.5*	*F28H7.3*	*F32B5.8*	*F41H10.7*	*F41H10.8*	*F42A10.6*
F44C4.3	*F46E10.1b*	*F53A9.8*	*F54F11.2*	*F57F4.4*	*F57F5.1*	*F58B3.1*	*F58G1.4*
H06I04.4a	*H22K11.1*	*K03A1.2*	*K07C6.4*	*K07H8.6a*	*K09D9.2*	*K09F5.3a*	*K11D2.2*
K12H4.7a	*M02D8.4a*	*M03A8.1*	*M04G12.2*	*M88.1*	*R02E12.6*	*R06C1.4*	*R07E3.1*
R09B5.6	*R09F10.1*	*R57.1a*	*T01D3.6a*	*T07C4.4*	*T10B5.7*	*T15B7.2*	*T20G5.2*
T21H3.1	*Y119D3B.21*	*Y22F5A.4*	*Y22F5A.5*	*Y39B6A.1*	*Y49E10.16*	*Y54F10AM.8*	*Y69F12A.2a*
ZK1320.2	*ZK896.7*						

Table 2.1 – List of 74 Intestine-specific Transcripts

We identified these 74 transcripts as being specifically expressed in the *C. elegans* intestine by comparing a SAGE library made from dissected intestine to a SAGE library made from whole adult worms.

Sequence	Gene	Distance from ATG	Reference
TTCTGATAAGGG	*vit-2* vitellogenin	-159	[37,38]
CATTGATAAGCT		-105	
AACTGATAGCAA	*ges-1* carboxylesterase	-1135	[39]
AACTGATAAGGG		-1123	
TACTGATAAGAA	*cpr-1* cysteine protease	-175	[40]
GATTGATAAGAC		-79	
AACTGATAAAAT	*mtl-1* metallothionein	-319	[33]
AACTGATAAAGG	*mtl-2* metallothionein	-305	
AGCTGATAACAG		-90	
TGATGATAAAGT	*gcs-1* glutamyl-cysteine synthetase	-116	[32]
TGTTGATAAGAT	*smd-1* S-adenosylmethionine decarboxylase	-874	[41]
CACTGATAACGA		-860	
GGTAGATAGAAC		-795	
AGGTGATAAGAT	*Y46G5A.19* spermidine synthase	-133	
TAGTGATAATGG		-119	
CAGTGATAATAG		-110	
AGTTGATAGTGA		-97	
TTGTGATAATGA	*spl-1* sphingosine-1-phosphate lyase	-320	[42]
AACTGATAAAAG	*pho-1* acid phosphatase	-122	[17]

Table 2.2 – Experimentally Validated ELT-2 Binding Sites

In order to compare sequences bound by the transcription factor ELT-2, we assembled this list of experimentally validated ELT-2 binding sites from the literature.

Gene	Locus	CHE-1 site	Distance upstream	Strand	Score
R03C1.3	*cog-1*	ATGAAGCCGTAGATAG	-3402	-	0.69
K01B6.1	*fozi-1*	AAGAAGCCTTAAAAGT	-775	+	0.72
K03E6.1	*lim-6* #1	TGGAAACCTTATGAGC	-130	-	0.64
K03E6.1	*lim-6* #2	GTGAAGCACCTTATAA	-110	+	0.57
C37E2.4	*ceh-36*	AAGAAGCCTTAGAACC	-430	+	0.72
C55B7.12	*che-1*	GTGAAGCCACAATTTT	-250	+	0.62
C32C4.6	*lsy-6*	CGGAAGCCGTAAAATA	-104	-	0.7
ZK652.5	*ceh-23*	TGGAAGCCAATTATTT	-862	-	0.63
AH6.1	*gcy-1* #1	TAGAAGCCGCAAAAAG	-280	-	0.67
AH6.1	*gcy-1* #2	ACGAAGCCACTTTTTA	-142	+	0.6
R134.1	*gcy-3*	TAGAAGCCGTGTTTTC	-133	+	0.63
ZK970.5	*gcy-4*	AAGAAGCCAATCATCT	-35	+	0.63
ZK970.6	*gcy-5*	AAGAAGCCCCCAAATG	-160	+	0.63
B0024.6	*gcy-6*	TAGAAGCCTACAAACA	-1463	+	0.63
F52E1.4	*gcy-7*	GTGAAACCTTATTTTT	-109	-	0.65
ZC412.2	*gcy-14*	ATGAAACCTTGCAATA	-57	-	0.65
C17F4.6	*gcy-19*	GAGAAGCCGTACAACT	-675	+	0.71
F21H7.9	*gcy-20*	AAGAAACCTTTCAATA	-60	-	0.66
T03D8.5	*gcy-22*	GGGAAGCCCTTCAAAT	-145	-	0.69
F48C11.3	*nlp-3*	TAGAAGCCCCTCACAA	-1292	-	0.61
F18E9.2	*nlp-7*	GTGAAACCCTGTTAAG	-1769	-	0.61
F07D3.2	*flp-6*	AAGAAGCCTTATTAGA	-1504	+	0.65
F33D4.3	*flp-13*	AGGAATCCCTACAAGA	-46	-	0.66
E01H11.3	*flp-20*	AAGAAACCTTATCTTT	-765	-	0.64
R13H7.2	*R13H7.2*	TGGAAGCCGTAGCTTT	-3696	+	0.64
F55E10.7	*F55E10.7*	TTGAAACCATAGACTA	-847	+	0.63
C36B7.7	*hen-1*	TGGAATCCTTAGATCC	-857	+	0.66

Table 2.3 – Experimentally Validated CHE-1 Binding Sites

In a series of experiments using GFP expression constructs, we assembled a series of 27 CHE-1 binding sites upstream of 25 ASE-specific genes.

	1	2	3	4	5	6	7	8	9	10
A	0.74	0.28	0.07	0	1.00	0	1.00	0.97	0.19	0.54
C	0.04	0.29	0.01	0	0	0	0	0	0.15	0.09
G	0.13	0.13	0.01	1.00	0	0	0	0.02	0.53	0.2
T	0.09	0.31	0.91	0	0	1.00	0	0.01	0.13	0.17

Table 2.4 – Position Frequency Matrix of Experimental and Putative ELT-2 Binding Sites

Experimentally validated ELT-2 binding sites were combined with the large cluster of motifs discovered in the upstream region of the 74 genes to form an estimate of the ELT-2 binding site. Base frequencies in the central 10 conserved bases were tabulated and normalized to make the sum of each column equal to one. This PFM was subsequently used to score matches to the pattern NNNGATARNN found in the upstream regions of *C. elegans* genes.

PFM	1	2	3	4	5	6	7	8	9	10	11	12
A	0	1.00	1.00	0.26	0	0.04	0.19	0.11	0.56	0.22	0.67	0.52
C	0	0	0	0	1.00	0.96	0.22	0.22	0.07	0.26	0.07	0.07
G	1.00	0	0	0.67	0	0	0.22	0	0.11	0.19	0.04	0
T	0	0	0	0.07	0	0	0.37	0.67	0.26	0.33	0.22	0.41

Table 2.5 – Position Frequency Matrix of Experimentally Validated CHE-1 Sites

We generated a position frequency matrix from the CHE-1 binding sites in order to search for similar sites in the upstream regions of various groups of genes.

2.7 Chapter 2 Figures

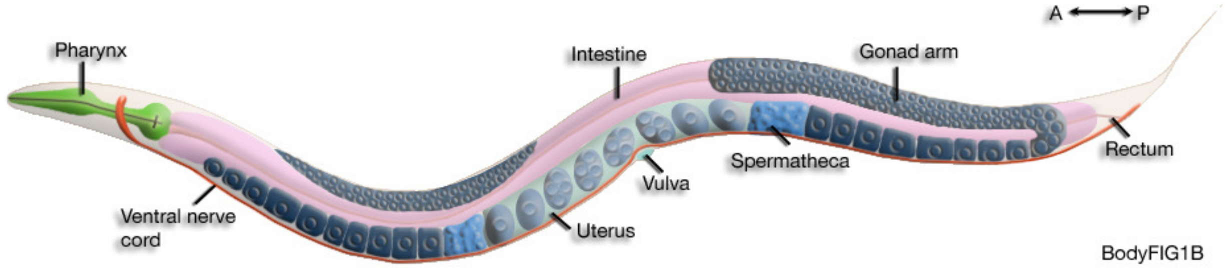


Figure 2.1 – Diagram of Basic Worm Anatomy

Diagram of three major systems of *C. elegans* hermaphrodite anatomy: The digestive system, which is made up of the pharynx, intestine, and rectum; The reproductive system, which is made up of two gonad arms, spermatheca, uterus, fertilized eggs, and vulva; and the nervous system, which is made up of head neurons, the ventral nerve cord, and the dorsal nerve cord (not shown). Image Copyright 2005 Wormatlas www.wormatlas.org [43]. Used with permission.

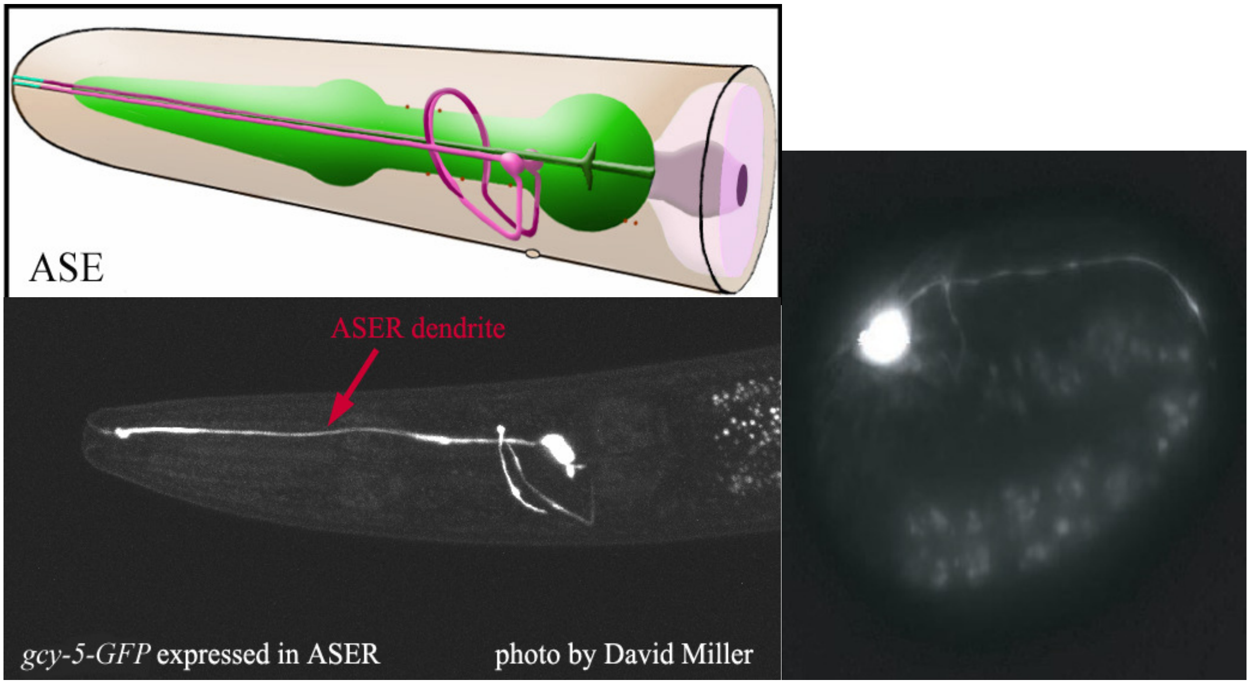


Figure 2.2 – Diagram and Image of the ASE Neuron in the *C. elegans* Head

Top Left: Diagram showing the locations of both ASE neurons. Chemoreceptors are exposed to the environment via cilia in the amphids near the mouth (cyan portion at left). The dendrites extend along the length of the head; the neurons also make contact with other sensory neurons and interneurons in the nerve ring (looped portion). Bottom Left: An adult *C. elegans* expressing the gcy-5::GFP construct. Because the ASER neuron is the only cell in *C. elegans* to express *gcy-5*, the construct will clearly distinguish this cell from all others. Top left and bottom left images Copyright 2005 Wormatlas www.wormatlas.org [44]. Used with permission. Right: A *C. elegans* embryo expressing the GFP construct in the ASER cell at the 3-fold stage. Image Copyright 2007, Genes and Development Online, www.genesdev.org, by Cold Spring Harbor Laboratory Press [2]. Used with permission.

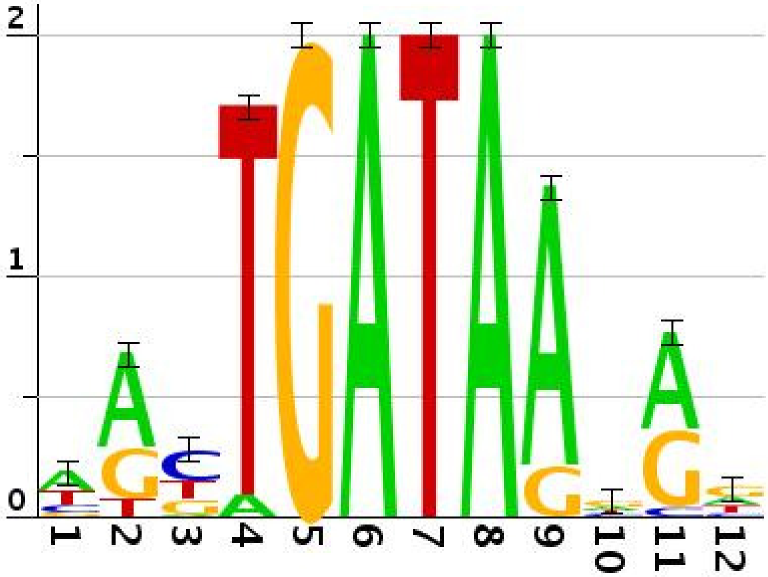


Figure 2.3 – Logo of Experimentally Validated ELT-2 Binding Sites

Logo showing the TGATAR motif that characterizes the ELT-2 binding sites from Table 2.2. Image substantially similar to an image from Developmental Biology, Volume 302, by McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA, Baillie DL, Kohara Y, Marra MA, Jones SJ, Moerman DG, and Robertson AG, "The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine", pp 627-645, Copyright 2007, with permission from Elsevier.

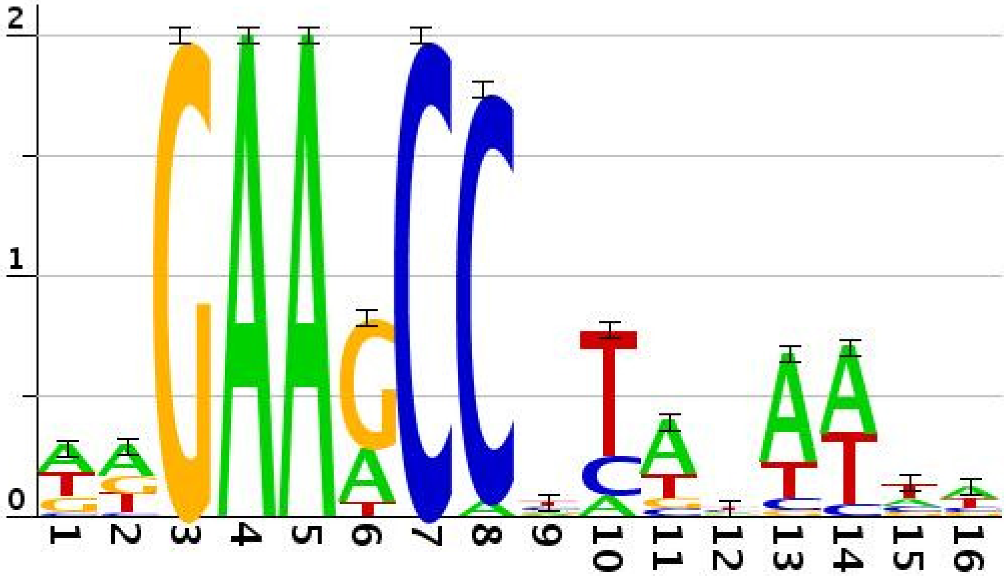


Figure 2.4 – Logo of Experimentally Validated CHE-1 Binding Sites

Logo of the 27 experimentally validated CHE-1 binding sites from Table 2.3. Note that the first half of the motif is characterized by a prominent GAARCC pattern while the second half is poorly conserved.

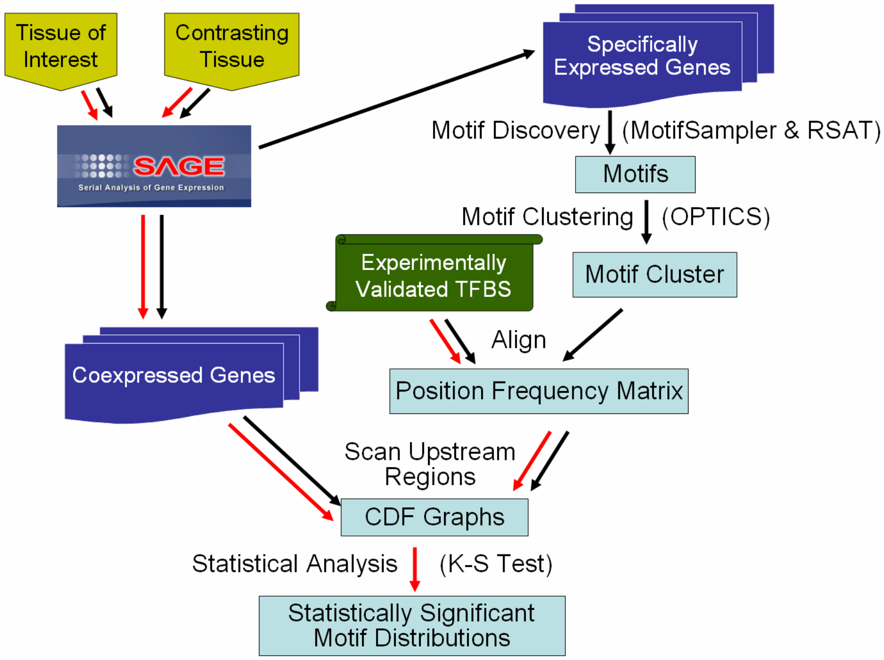


Figure 2.5 – Flowchart of Approach

Analysis steps carried out for the Intestinal study are indicated with black arrows while steps carried out for the ASE Neuron study are indicated with red arrows. In both cases, we used two contrasting SAGE libraries, one made from the tissue of interest (intestine and ASER neurons respectively) and the other made from a contrasting tissue (whole worm and AFD neurons respectively), to generate sets of genes that were expressed more strongly in one tissue than the other. For the Intestinal study, we also used the SAGE libraries to generate a list of 74 genes that were specifically expressed in the intestine. We then used the motif discovery programs MotifSampler and RSAT to find motifs in the upstream regions of these genes. The discovered motifs were formed into clusters using the OPTICS clustering algorithm, and members of the largest cluster, which were all variations on the pattern NNNGATARNN, were combined with experimentally validated ELT-2 binding sites to form a Position Frequency Matrix (PFM). For the ASE Neuron study, motif discovery was not performed and only experimentally validated CHE-1 binding sites were used to generate the PFM because there were many more experimentally validated sites. In both studies, the PFM was used to scan the upstream regions of various sets of genes, and the cumulative distribution functions of the highest-scoring results for the gene sets were graphed. Finally, for the ASE Neuron study only, the differences between the distribution functions were smaller and the Kolmogorov-Smirnov test was used to identify which pairs of gene sets had statistically significant differences in the distribution of high-scoring sites.

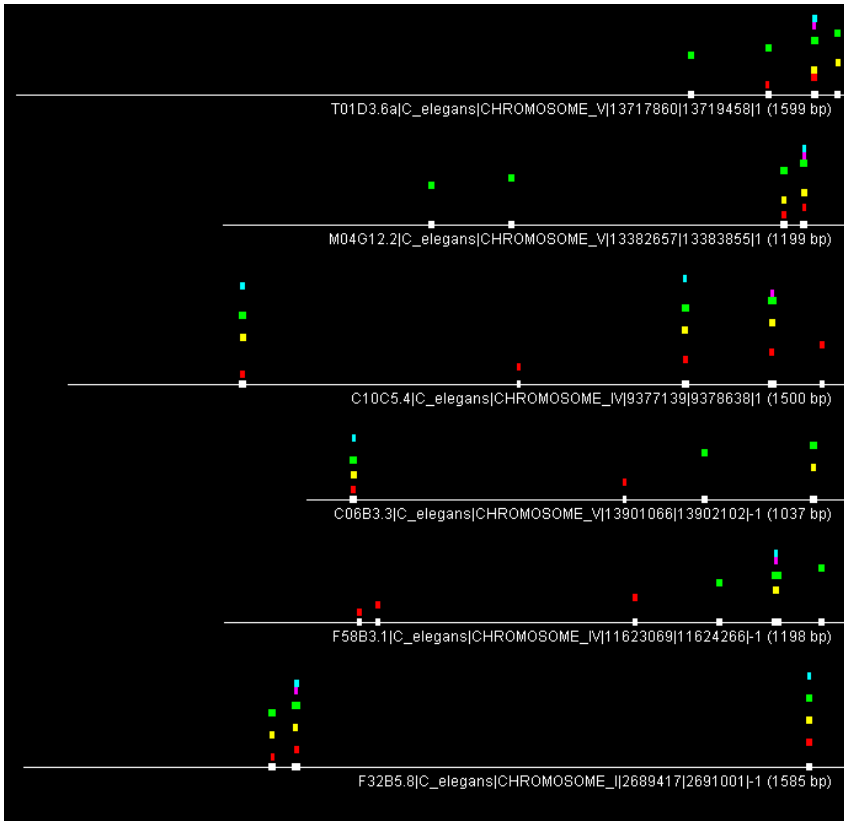


Figure 2.6 – Motif Discovery Example

Diagram showing findings for six genes as an example. RSAT results (width 8) are in red; MotifSampler results: Width 6 are magenta, Width 8 are cyan, Width 10 are yellow, Width 12 are green. The combined result, which we used for the rest of the analysis, is shown as white boxes. Note that for some genes the RSAT and MotifSampler results agree perfectly, and for some genes they disagree perfectly, but for most genes they agree on some instances and disagree on others.

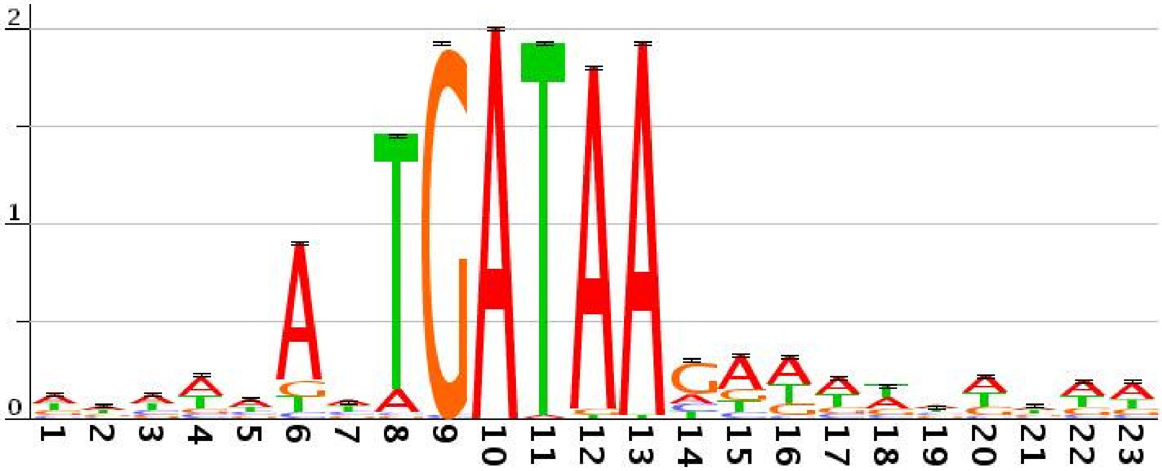


Figure 2.7 – Logo of Largest Cluster of Sequences from Motif Discovery in *C. elegans*

Motifs obtained from the motif discovery process were aligned using ClustalW and clustered using OPTICS. The sequence logo of the largest cluster, consisting of 111 sequences from 59 of the genes, is shown.

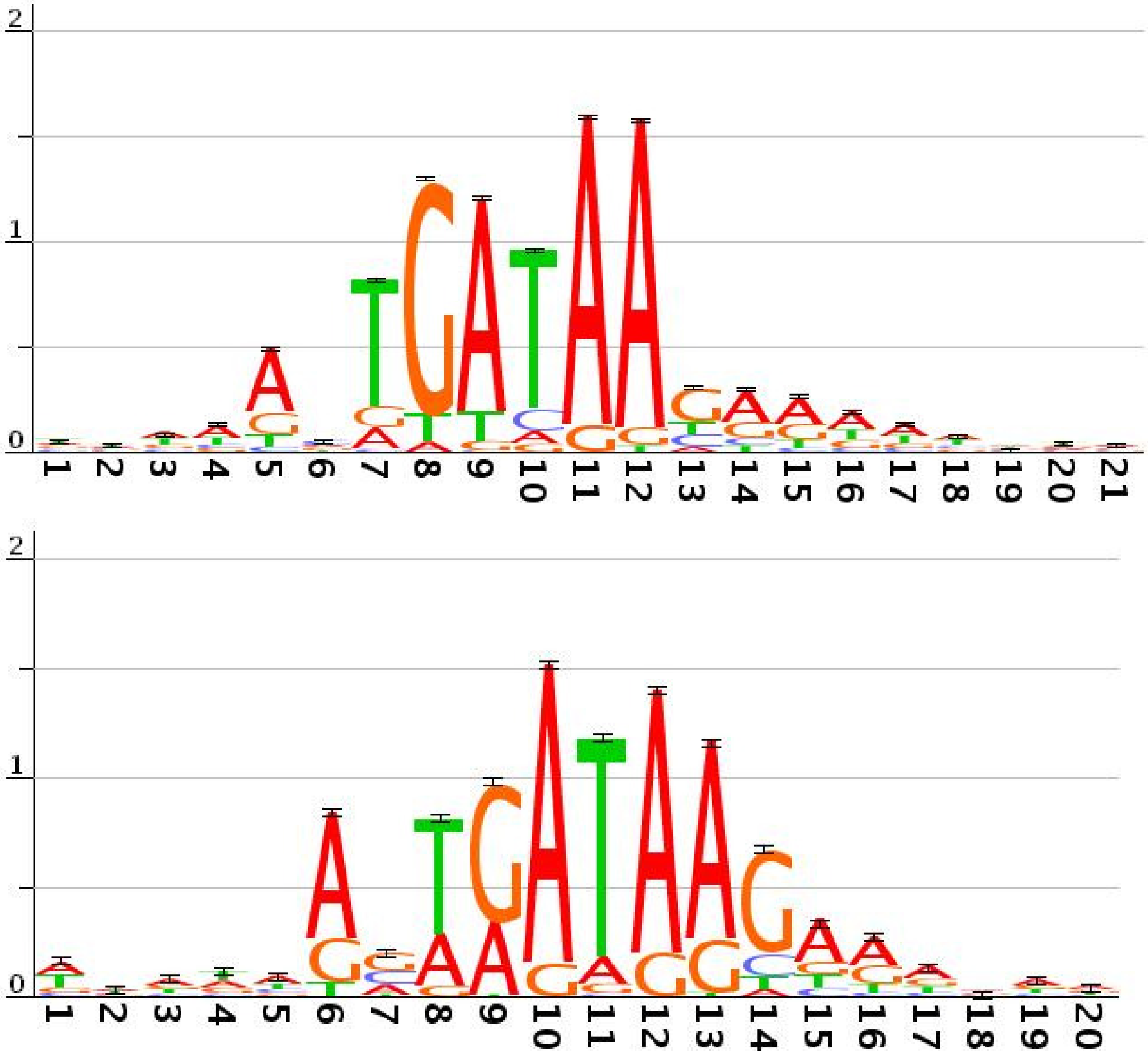


Figure 2.8 – Logo of the Largest Cluster of Motifs from *C. briggsae* and *C. remanei*

Motif discovery, alignment, and clustering from the upstream regions of 57 *C. briggsae* (top) and 39 *C. remanei* (bottom) orthologues of intestine-specific genes yielded a result very similar to what was seen in *C. elegans*.

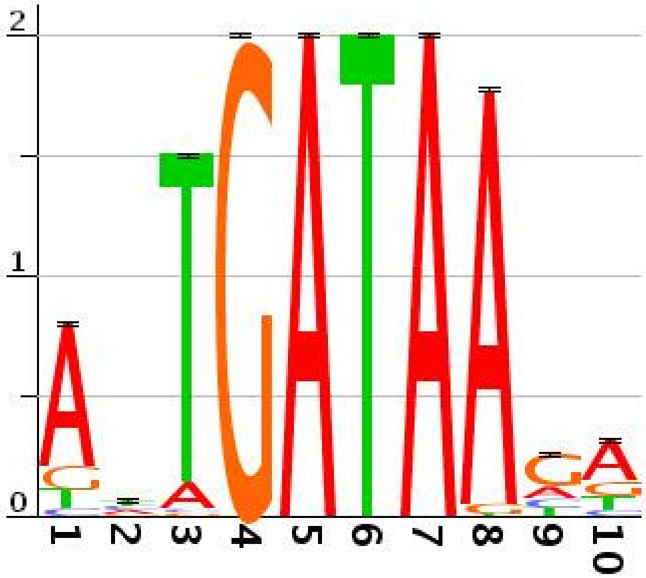


Figure 2.9 – Logo of Experimental and Putative ELT-2 Binding Sites

This sequence logo shows the conservation of the sequences used to form the PFM in Table 2.4. Note that the bases in positions four to seven are invariant, forming the core of the GATA-factor binding site. Image substantially similar to an image from Developmental Biology, Volume 302, by McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA, Baillie DL, Kohara Y, Marra MA, Jones SJ, Moerman DG, and Robertson AG, "The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine", pp 627-645, Copyright 2007, with permission from Elsevier.

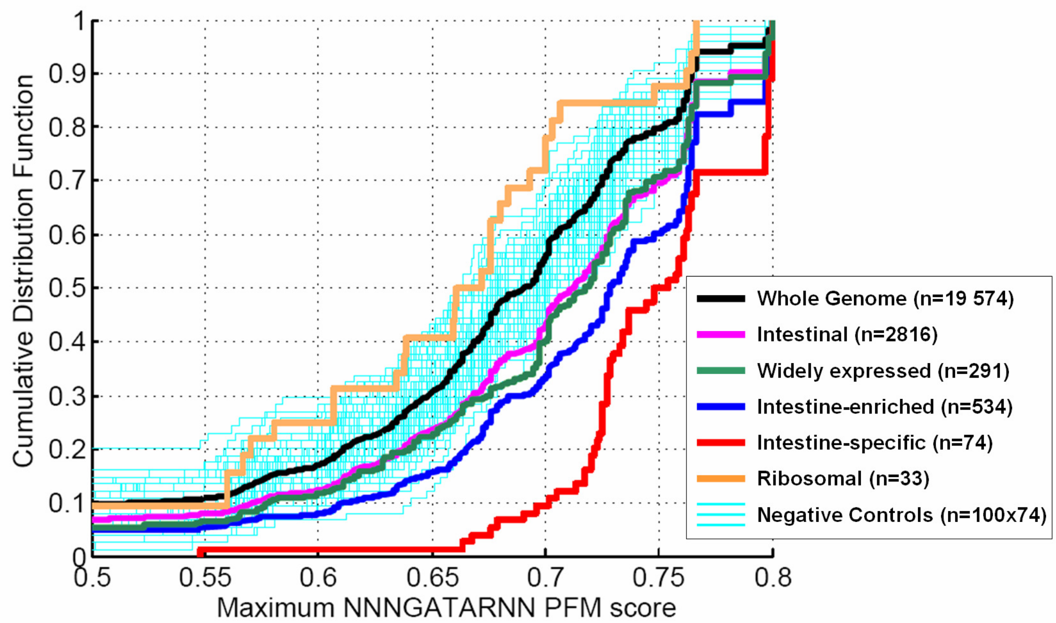


Figure 2.10 – Cumulative Distribution Function of Maximum Intestinal PFM Score

Lines are colour coded as follows: Black = all protein-coding promoters in the *C. elegans* genome; Magenta = promoters from genes expressed in the intestine (intestine tag count > 1); Green = promoters from widely-expressed genes (Intestine tag count ≥ 9; somatic tag count >0; 0.67 ≤ I/S tag ratio ≤ 1.5); Blue = promoters from intestine enriched genes (intestine tag count ≥ 9; somatic tag count > 0; I/S tag ratio ≥ 2); Red = The 74 highly expressed intestine specific promoters (Table 2.1) used in the computational analysis; Orange = promoters from ribosomal protein genes expressed in the adult intestine (somatic tag count > 0; I/S tag ratio ≥ 2); Cyan = 100 independent random samplings of 74 promoters from the entire genome. Image substantially similar to an image from Developmental Biology, Volume 302, by McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA, Baillie DL, Kohara Y, Marra MA, Jones SJ, Moerman DG, and Robertson AG, “The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine”, pp 627-645, Copyright 2007, with permission from Elsevier.

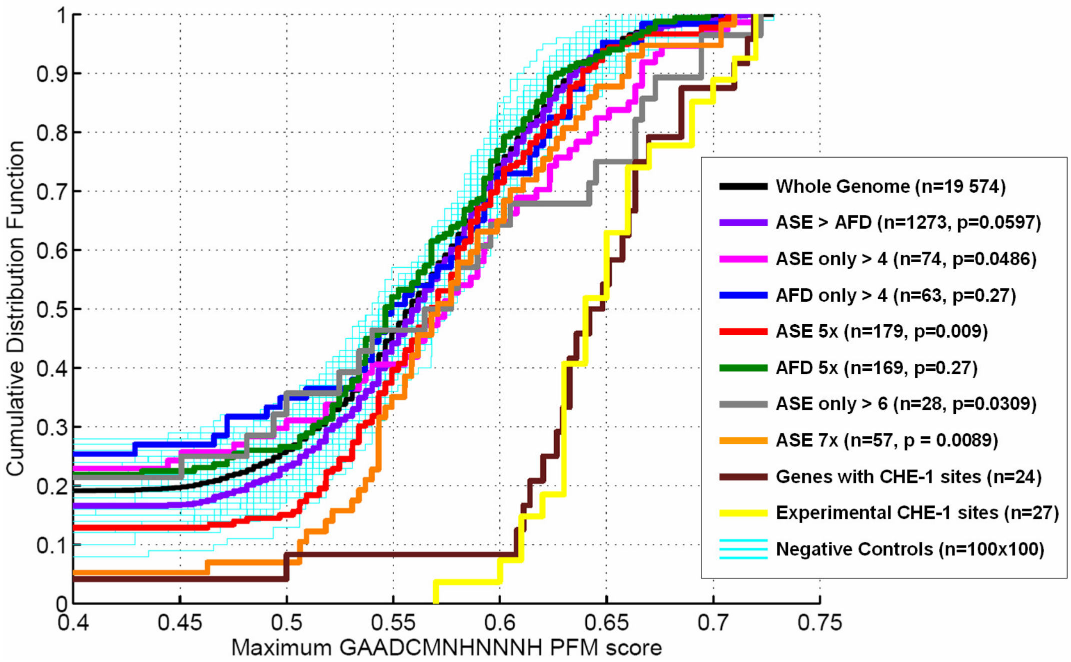


Figure 2.11 – Cumulative Distribution Function of Maximum ASE Neuron PFM Score

Lines are colour coded as follows: Black = all promoters in the *C. elegans* genome; Purple = promoters from the ASE>AFD gene set; Magenta = promoters from genes with a tag count > 4 in the ASE library and a count of 0 in the AFD library; Blue = promoters from genes with a tag count > 4 in the AFD library and a count of 0 in the ASE library; Red = promoters from genes with an ASE:AFD tag ratio ≥ 5; Green = promoters from genes with an AFD:ASE tag ratio ≥ 5; Grey = promoters from genes with a tag count > 6 in the ASE library and a count of 0 in the AFD library; Orange = promoters from genes with an ASE:AFD tag ratio ≥ 7; Brown = promoters from protein-coding genes with experimentally validated CHE-1 sites in Table 2.3; Yellow = Experimentally validated CHE-1 binding sites from Table 2.3; Cyan = 100 independent random samplings of 100 promoters from the entire genome. Image substantially similar to an image from Etchberger *et al.*, Copyright 2007, Genes and Development Online, www.genesdev.org, by Cold Spring Harbor Laboratory Press [2]. Used with permission.

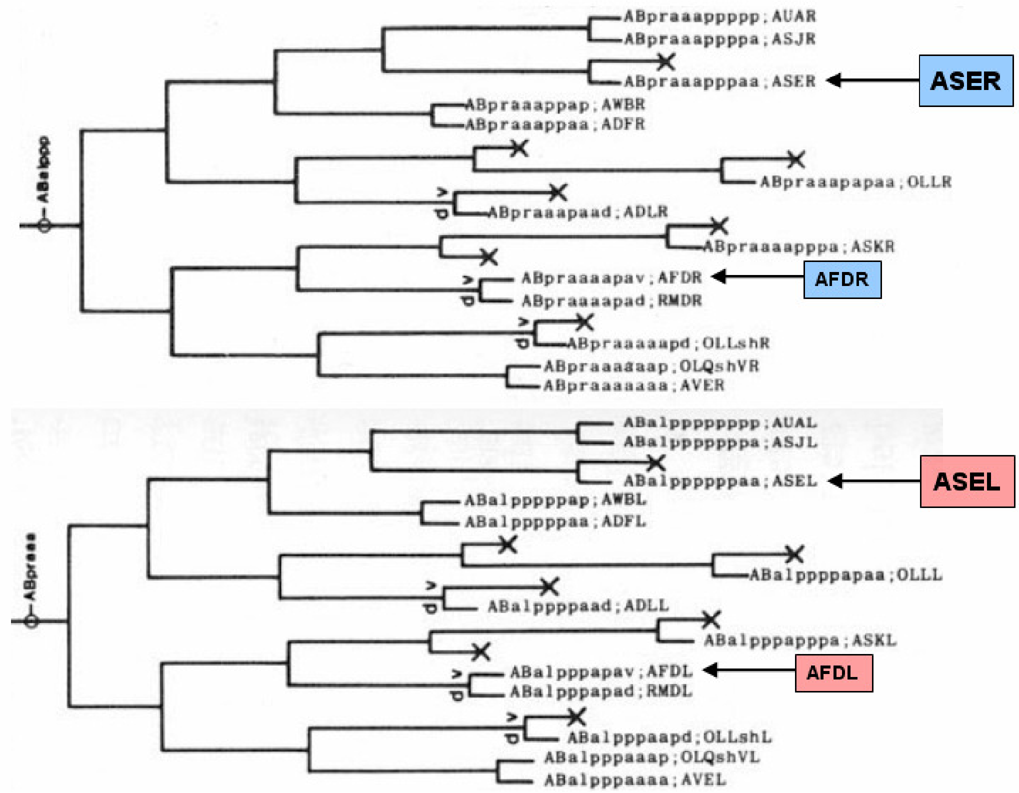


Figure 2.12 – Embryonic Cell Lineages of ASE and AFD Neurons

Excerpts of the embryonic cell lineage as determined by John Sulston showing the origins and related cells of ASER, AFDR, ASEL, and AFDL. Top: The ASER and AFDR neurons both originate from a lineage that produces a large number of other neurons on the right side of the worm. Bottom: ASEL and AFDL originate from a corresponding lineage for neurons on the left side. The two lineages separate at the four-cell stage of embryogenesis: the right-sided neurons originate from the ABa cell and the left-sided neurons originate from the ABp cell (not shown). Image reprinted from Developmental Biology, Volume 100, by Sulston JE, Schierenberg E, White JG, and Thomson JN, “The embryonic cell lineage of the nematode *Caenorhabditis elegans*”, pp 64-119, Copyright 1983, with permission from Elsevier.

2.8 References

1. McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ *et al.* (2007) The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. Dev Biol 302: 627-645.
2. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ *et al.* (2007) The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Genes Dev 21: 1653-1674.
3. Deppe U, Schierenberg E, Cole T, Krieg C, Schmitt D *et al.* (1978) Cell lineages of the embryo of the nematode Caenorhabditis elegans. Proc Natl Acad Sci U S A 75: 376-380.
4. Sulston JE, Schierenberg E, White JG, Thomson JN (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol 100: 64-119.
5. Altun ZF and Hall DH. (2005) Alimentary System - (Part II) The Intestine. In Wormatlas http://www.wormatlas.org/handbook/alimentary/alimentary2.htm
6. McGhee JD (2007) The C. elegans intestine. WormBook 1-36.
7. Hobert O (2005) Specification of the nervous system. WormBook 1-19.
8. Hall DH and Russell RL (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.
9. Hart AC, ed. (2005) Behavior. WormBook 1-36.
10. Altun ZF and Hall DH. (2006) *Caenorhabditis elegans* as a Genetic Organism. In Wormatlas http://www.wormatlas.org/handbook/anatomyintro/anatomyintro.htm
11. Inglis PN, Ou G, Leroux MR, Scholey JM (2007) The sensory cilia of Caenorhabditis elegans. WormBook 1-22.
12. Johnston RJ,Jr, Chang S, Etchberger JF, Ortiz CO, Hobert O (2005) MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision. Proc Natl Acad Sci U S A 102: 12449-12454.
13. Bargmann CI and Horvitz HR (1991) Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in C. elegans. Neuron 7: 729-742.
14. Uchida O, Nakano H, Koga M, Ohshima Y (2003) The C. elegans che-1 gene encodes a zinc finger transcription factor required for specification of the ASE chemosensory neurons. Development 130: 1215-1224.
15. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487.
16. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL *et al.* (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68: 159-169.
17. Fukushige T, Goszczynski B, Yan J, McGhee JD (2005) Transcriptional control and patterning of the pho-1 gene, an essential acid phosphatase expressed in the C. elegans intestine. Dev Biol 279: 446-461.
18. Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.
19. Inada H, Ito H, Satterlee J, Sengupta P, Matsumoto K *et al.* (2006) Identification of guanylyl cyclases that function in thermosensory neurons of Caenorhabditis elegans. Genetics 172: 2239-2252.
20. Benos PV, Lapedes AS, Stormo GD (2002) Probabilistic code for DNA recognition by proteins of the EGR family. J Mol Biol 323: 701-727.
21. Moses K and Rubin GM (1991) Glass encodes a site-specific DNA-binding protein that is regulated in response to positional signals in the developing Drosophila eye. Genes Dev 5: 583-593.

22. Yan H, Canon J, Banerjee U (2003) A transcriptional chain linking eye specification to terminal determination of cone cells in the Drosophila eye. Dev Biol 263: 323-329.
23. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9: 447-64.
24. van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res 31: 3593-6.
25. Gaudet J, Muttumu S, Horner M, Mango SE (2004) Whole-genome analysis of temporal gene expression during foregut development. PLoS Biol 2: e352.
26. GuhaThakurta D, Schriefer LA, Waterston RH, Stormo GD (2004) Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. Genome Res 14: 2457-2468.
27. Kent WJ and Zahler AM (2000) Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res 10: 1115-25.
28. Ankerst, M., Breunig, M., Kriegel, H.-P., Sander,J.(1999) Ordering points to identify the clustering structure. ACM SIGMOD 1(1): 49-60.
29. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-80.
30. Maduro MF, Hill RJ, Heid PJ, Newman-Smith ED, Zhu J *et al.* (2005) Genetic redundancy in endoderm specification within the genus Caenorhabditis. Dev Biol 284: 509-522.
31. Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-87.
32. An JH and Blackwell TK (2003) SKN-1 links C. elegans mesendodermal specification to a conserved oxidative stress response. Genes Dev 17: 1882-1893.
33. Moilanen LH, Fukushige T, Freedman JH (1999) Regulation of metallothionein gene transcription. Identification of upstream regulatory elements and transcription factors responsible for cell-specific expression of the metallothionein genes from Caenorhabditis elegans. J Biol Chem 274: 29655-29665.
34. Sandelin A and Wasserman WW (2005) Prediction of nuclear hormone receptor response elements. Mol Endocrinol 19: 595-606.
35. Pierce-Shimomura JT, Faumont S, Gaston MR, Pearson BJ, Lockery SR (2001) The homeobox gene lim-6 is required for distinct chemosensory representations in C. elegans. Nature 410: 694-698.
36. Chen N, Mah A, Blacque OE, Chu J, Phgora K *et al.* (2006) Identification of ciliary and ciliopathy genes in Caenorhabditis elegans through comparative genomics. Genome Biol 7: R126.
37. MacMorris M, Broverman S, Greenspoon S, Lea K, Madej C *et al.* (1992) Regulation of vitellogenin gene expression in transgenic Caenorhabditis elegans: short sequences required for activation of the vit-2 promoter. Mol Cell Biol 12: 1652-1662.
38. MacMorris M, Spieth J, Madej C, Lea K, Blumenthal T (1994) Analysis of the VPE sequences in the Caenorhabditis elegans vit-2 promoter with extrachromosomal tandem array-containing transgenic strains. Mol Cell Biol 14: 484-491.
39. Egan CR, Chung MA, Allen FL, Heschl MF, Van Buskirk CL *et al.* (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.
40. Britton C, McKerrow JH, Johnstone IL (1998) Regulation of the Caenorhabditis elegans gut cysteine protease gene cpr-1: requirement for GATA motifs. J Mol Biol 283: 15-27.
41. Luersen K, Eschbach ML, Liebau E, Walter RD (2004) Functional GATA- and initiator-like-elements exhibit a similar arrangement in the promoters of Caenorhabditis elegans polyamine synthesis enzymes. Biol Chem 385: 711-721.

42. Oskouian B, Mendel J, Shocron E, Lee MA,Jr, Fyrst H *et al.* (2005) Regulation of sphingosine-1-phosphate lyase gene expression by members of the GATA family of transcription factors. J Biol Chem 280: 18403-18410.
43. Altun ZF and Hall DH. (2005) The Nematode Body Shape. In Wormatlas http://www.wormatlas.org/handbook/bodyshape.htm
44. Altun ZF and Hall DH. (2005) ASE Neurons. In Wormatlas http://www.wormatlas.org/neurons.htm/ase.htm

3 *C. elegans* cisRED: A Catalogue of Conserved Genomic Elements[2]

3.1 Introduction

The binding of transcription factors (TFs) to DNA sequences upstream of a gene is an important element in transcriptional control [1]. The genome of the nematode *C. elegans* is well characterized and almost all of its genes have been identified [2], including 664 genes predicted to encode TFs [3]. However, binding sites have been identified for less than 50 of these TFs, and transcriptional regulation is understood for only a few genes. Because regulatory elements are shared among the upstream regions of orthologous [4,5] and coexpressed [6,7] genes, computational methods involving DNA sequence motif discovery among upstream regions of putative coregulated (orthologous or coexpressed) genes have been used to direct laboratory experiments such as reporter gene and gel shift assays [5,8]. Recently, the pace of genome sequence generation has increased and the assembled sequences of eight nematode species have become publicly available. Here, we take advantage of this information and attempt to predict regulatory elements in upstream regions of *C. elegans* genes by comparing these regions to orthologous regions in other nematode genomes. We hypothesized that most regulatory elements are conserved between many of the eight species, and conversely, that many conserved promoter elements have regulatory function.

To find novel regulatory elements in the *C. elegans* genome using a comparative genomics approach, we used eight sequenced nematode genomes that were available from either the WormBase [2] or Washington University Genome Sequence Center public FTP servers (Supplementary Table 1). These included the genome sequences or assemblies of *C. elegans* [9], *C. briggsae* [10], *C. remanei* (unpublished), *C. brenneri* [11], *C. japonica* (unpublished), *Pristionchus pacificus* [12], *Brugia malayi* [13], and *Trichinella spiralis* [14].

The first five of these species are in the same genus as *C. elegans* [15] (Figure 3.1). *C. elegans* diverged from the other species in genus *Caenorhabditis* between 18 and 100 million years ago [10,16]. *P. pacificus* is similar to *Caenorhabditis* species in that it is also a free-living soil bacteriovore, and is grouped in the same clade; *C. elegans* and *P. pacificus* diverged between 280 and 430 million years ago [12]. *B. malayi* and *T. spiralis* are mammalian parasites from different clades [17], and are therefore much more remotely related. *C. elegans* and *B. malayi*

[2] A version of this chapter has been published. Monica C. Sleumer, Mikhail Bilenky, An He, Gordon Robertson, Nina Thiessen, and Steven J. M. Jones ***Caenorhabditis elegans*** **cisRED: a catalogue of conserved genomic elements** (2009) Nucleic Acids Research 37(4):1323-1334, nar.oxfordjournals.org. Copyright 2009 Oxford University Press.

diverged between 350 and 540 million years ago [12], while *C. elegans* and *T. spiralis* diverged more than 600 million years ago [14].

Of the eight nematode genomes, only *C. elegans* has been extensively characterized in terms of gene location, expression, and function. Given this, we first identified orthologues for *C. elegans* protein-coding genes in the other seven genomes using WABA (Figure 3.2) [18]. Although genes have been predicted for some of the species, and orthologues from *C. elegans* to *C. briggsae* and *C. remanei* have been inferred, we chose to use a single consistent orthologue prediction method for all species. We included alternative transcripts for *C. elegans* genes because such transcripts frequently have different translation start sites (ATG) and transcripts with the same ATG can have different predicted orthologues if the coding exons vary widely.

We then assembled sets of orthologous upstream sequence regions (Figure 3.2). To do this, we pooled the upstream region of each *C. elegans* transcript with that of its predicted orthologues, extending each upstream region to the next protein-coding sequence, to a maximum of 1500 base pairs (bp). We used the Gibbs sampler MotifSampler [19] to find conserved DNA sequence motifs in each set of upstream region sequences. All motifs were loaded into the *C. elegans* cisRED database [20] and are publicly available via the database web interface at www.cisred.org. We used 44 experimentally validated transcription factor binding sites (TFBSs) from ORegAnno [21], found in 28 of the upstream regions, to validate the motif discovery process. Lastly, we compared motif sequences to TF binding sequences from TRANSFAC [22], JASPAR [23], and ORegAnno, and annotated a motif as similar to a binding sequence if the comparison was statistically significant.

3.2 Results

3.2.1 Orthologue Identification

For each of the 23,212 *C. elegans* chromosomal protein-coding transcripts, we used the WABA algorithm [18] to identify putative orthologues in the other seven genomes. WABA is similar to BLAST and was originally designed for use in nematodes [10,24]. We found WABA to be particularly useful for our purposes because it finds putative orthologues for protein-coding DNA sequences from an annotated genome to a newly assembled, unannotated genome without intermediate gene prediction and translation steps.

WABA and InParanoid results were concordant. In order to determine whether WABA results were reliable compared to protein-level orthologue determination, we compared its output to the InParanoid database [25]. We found that InParanoid identified 12,197 one-to-one orthologues between *C. elegans* and *C. briggsae* genes, while WABA identified single orthologues for 12,326 *C. elegans* transcripts (Figure 3.3). Of these 12,326, InParanoid also had identified single orthologues for 11,231 (91% of 12,326 and 92% of 12,197). Of the 11,231 *C. elegans* transcripts with both a single WABA orthologue and a single InParanoid orthologue, the WABA orthologue overlapped the InParanoid orthologue for 11,104 (98.9%), and the start site of the WABA orthologue was within 750 bp of that of the InParanoid orthologue for 9645 (86%).

C. brenneri **had two matches for many** ***C. elegans*** **transcripts.** All four species from genus *Caenorhabditis* had at least one match for 14,000 to 18,000 of the *C. elegans* transcripts (Figure 3.3). *C. briggsae* and *C. remanei* both had single matches for about 12,000 *C. elegans* transcripts and two matches for approximately 3000 additional transcripts. However, for *C. brenneri*, a disproportionately small number of *C. elegans* WormPep sequences had one match and a large number had two matches. The result was that far fewer *C. elegans* transcripts had suitable orthologues in *C. brenneri* (< 4500) than in the other two *Caenorhabditis* species (> 6000), even though all three species are the same evolutionary distance from *C. elegans*. As expected, the three more distant nematode species (*P. pacificus*, *B. malayi*, *T. spiralis*) had far fewer WABA-predicted orthologues than the more closely related nematodes.

Because the analysis described in this paper involved regions directly upstream of ATGs, it was important to accurately identify the N-terminal of each orthologue. Therefore, only high-quality orthologues, i.e. single WABA matches that started at the ATG of the *C. elegans* transcript, were used for the next step of the analysis.

3.2.2 Orthologous Upstream Sequence Regions

Orthologous upstream sequence region sets were formed by pooling the upstream region of each *C. elegans* transcript with that of its orthologues from the other genomes. Only transcripts with at least three out of a possible seven high-quality orthologues were retained. The resulting collection contained upstream sets for 3847 *C. elegans* transcripts, but was somewhat redundant due to both transcripts from the same gene that shared the same ATG and transcripts on bidirectional promoters that shared the same upstream region; 3544 different transcript upstream regions and 3458 genes were represented. Taking orthologous sequences into account, the collection contained 3551 unique upstream sets. WABA identified a unique region of each unannotated genome as an orthologue 96% of the time. Only 141 transcripts had orthologues that overlapped those of another transcript. These may be a result of a gene duplication event that occurred in *C. elegans* after it diverged from the other species.

Bidirectional promoters were highly conserved among nematodes. We identified 132 *C. elegans* bidirectional promoters shorter than 1500 bp, of which 25 (19%) were perfectly conserved among all species for which orthologues were found and another 89 (67%) were conserved among orthologues from other species in genus *Caenorhabditis*. Only 10 (8%) bidirectional promoters were not conserved in any of the species. We also noted that five (4%) of the transcript pairs on bidirectional promoters had similar or identical protein-coding sequences and as a result had the same orthologues.

Most transcripts only had orthologues in other species of genus *Caenorhabditis*; only 14% had orthologues in *P. pacificus*, *B. malayi*, or *T. spiralis*. There were 1027 (27%) *C. elegans* transcripts with orthologues in all four of the other *Caenorhabditis* species, and another 2298 (60%) transcripts had orthologues in three out of four of these species (Figure 3.4). Only 202 (5%) transcripts had orthologues in *P. pacificus* as well as in some *Caenorhabditis* species, 188 (5%) transcripts had orthologues from at least one of the two parasitic nematodes but not *P. pacificus*, and 116 (3%) transcripts had orthologues from both *P. pacificus* and a parasitic nematode. Only three transcripts had orthologues from *P. pacificus* and both parasitic nematodes but not from any species in *Caenorhabditis*. Finally, 13 transcripts had orthologues in all seven nematode species: *rpl-2* (*B0250.1*), *cyn-10* (*B0252.4b*), *rps-13* (*C16A3.9*), *phi-18* (*C37C3.2* transcripts *b&c*), *D1054.14*, *rps-9* (*F40F8.10*), *rpn-6* (*F57B9.10b*), *T10C6.5*, *cdc-37* (*W08F4.8a*), *W09G12.5* (now known as *F38A1.8*), *rab-30* (*Y45F3A.2*), and *aps-3* (*Y48G8AL.14*).

Chromosomes III and X were overrepresented among the transcripts in the set, while Chromosomes IV and V were underrepresented (Pearson χ^2 p-value < 10^{-15}). In contrast, the proportion of transcripts on Chromosomes I and II was not significantly different (Figure 3.5).

3.2.3 Motif Discovery

A multi-species high-order Markov background model improved MotifSampler's specificity. MotifSampler can use a high-order Markov background model to reduce the probability that it will return unmasked repeats and other low-complexity sequences as a motif [26]. This was important for nematode genomes because they are 57-70% AT and contain much low-complexity sequence.

Extensive testing was done to determine settings for MotifSampler parameters that maximized the sensitivity while minimizing the total number of motifs. We found that the sensitivity was >80% when we retained motifs with MotifSampler scores above the 70th percentile but decreased rapidly for score thresholds above the 80th percentile. The coverage (proportion of bases covered by at least one motif) decreased linearly as we increased the motif score threshold from the 50th to the 90th percentile. Therefore, we retained only the top 30% of motifs found by MotifSampler.

A substantial number of motifs were very wide. Of the total of 158,017 motifs found, 14 bp motifs were the most common of the five widths (Figure 3.6). After overlapping motifs were merged, the distribution of motif widths developed a long tail: many of the motifs were much wider than 14 bp, nearly 4000 motifs were $\geq$ 30 bp wide, and the widest motif was 212 bp.

Most motifs were found in all sequences of the orthologous upstream sequence region set. The majority of the upstream sequence region sets consisted of *C. elegans* and three or four sequences from other *Caenorhabditis* species (Figure 3.4). The motif discovery algorithm found 84% of motifs in all species of the sequence set, with the result that most motifs had a species depth (i.e. the number of species in which the motif was found) of four or five, including *C. elegans*. Four percent of motifs had a depth less than four, 59% of motifs had a depth of four, 33% had a depth of five, and 4% had a depth greater than five. All but 20 of the motifs had a depth of at least three. Motifs that were not found in all sequences came from upstream sequence sets in which one or more of the sequences was very different from the others. For example, the motifs were not found on a sequence from one of the more distant species or on a sequence that was highly repetitive.

The conserved proportion of upstream regions varied widely. Of all unmasked bases of *C. elegans* upstream regions, 45% were covered by at least one motif. The interquartile range of coverage of upstream regions was 36% to 58%, while a few upstream regions were nearly completely covered with motifs and other upstream regions were only 8% covered. There was a weak negative correlation ($r = -0.43$) between coverage and upstream length: shorter upstream sequences tended to have higher coverage (i.e. be more highly conserved). The spatial distribution of motifs across the upstream regions was uniform. No significant difference was seen between the distribution of motifs with respect to the ATG and the distribution of motifs with respect to the opposite end of the sequence (KS test, $p > 0.2$).

3.2.4 Validation

Discovered motifs were compared to experimentally validated TFBSs from the literature to gauge the success of the motif discovery process. For the 44 experimentally validated sites in the upstream regions under examination, 36 (82%) overlapped with motifs by at least 50% of the TFBS width, and 29 (66%) overlapped a motif completely. A complete list of experimentally validated sites and all cisRED motifs that overlapped them is shown in Supplementary Table 2.

For example, the following sites were found: the PHA-4 site near *tph-1* (*ZK1290.2b*) [27] (Figure 3.7 A), a DAF-12 site near *lit-1* (*W06F12.1c*) [28] (Figure 3.7 B), and an 'Early-2' motif near *K07C11.4* described by Gaudet *et al.* [4] (Figure 3.7 C). Of the eight known sites that were not found, seven were poorly conserved and one was a low-complexity PHA-4 site.

Motif p-values and information content were uncorrelated with motif function. We assigned a preliminary score to each motif using a simplified version of the scoring function described by Robertson *et al.* [20] in an attempt to evaluate its significance with respect to gene regulation. This score measured two parameters: depth of the motif (relative to the depth of the input set, which was from four to eight), and the average conservation of the bases (weighted by evolutionary distance, with more distant species weighted more heavily). The width of the motif was not included in the scoring function because experimentally validated TFBSs are as narrow as six bp and as wide as 16 bp. Each motif was then assigned a p-value indicating its rank in the distribution of scores of all 158,017 motifs. However, we found no relationship between the p-values and the functionality of the motifs; motifs overlapping experimentally validated sites were as likely to have a high p-value as a low p-value.

Motif information content (IC; a measure of the degree of conservation [29]) ranged from 0.7 bits to a perfectly conserved two bits with an the interquartile range of 1.45 to 1.75 (Figure 3.8). As was the case for the scoring function, IC was not useful in discriminating motifs that overlapped TFBSs; we observed no difference in the distribution of average IC between motifs overlapping experimentally validated sites and all motifs.

Functional regulatory elements were not the most highly conserved portions of the upstream regions. For example, we found 20 motifs in the 371 bp upstream region of *xbx-1* (*F02D8.3*) and its orthologues in *C. briggsae*, *C. remanei*, *C. brenneri*, and *C. japonica*, resulting in a coverage of 62% (Figure 3.9). This upstream region also contained an experimentally validated DAF-19 site [30], which was found by our method. However, five of the other motifs were more strongly conserved than the DAF-19 site (indicated by consensus sequence logos [31]; average IC also shown for each).

3.2.5 Annotation to Reveal Similarity to Known TFBSs

Five percent of the motifs were similar to TFBSs previously characterized in *C. elegans*. Motifs for which the *C. elegans* sequence displayed some similarity to one of 13 sets of TFBSs in *C. elegans* were identified and assigned a p-value indicating the significance of the similarity. We found that 36 of the motifs that overlapped experimentally validated sites by at least five bp could be annotated using this procedure. These could be separated into two groups:

20 motifs had very significant annotation p-values of < 0.00015, and the other 16 had less significant annotation p-values (p > 0.0009). Given this, the stringent threshold of 0.00015 was used for the ORegAnno binding sequence annotations. Four of the TFs had no annotated motifs below this threshold; sequences that were the same as or similar to these TFBSs appeared frequently enough among the non-conserved parts of the upstream regions that they could not be applied to the motifs with confidence. The TFs that were not annotated successfully were: PHA-4, DAF-12, the 'Heat Shock Element' described by GuhaThakurta *et al.* [32], and the 'Late-2' element described by Gaudet *et al.* [4]. A total of 7650 TF-motif combinations were annotated, representing 7449 different motifs; several motifs were annotated as similar to more than one TFBS. The most commonly annotated TFBS was DAF-19: 1305 motifs were annotated as similar to a DAF-19 site (Supplementary Table 3).

Eleven percent of the motifs were similar to TFBSs from TRANSFAC; 15% were similar to TFBSs from JASPAR. In order to determine whether any of the motifs were similar to binding sequences identified in species other than *C. elegans*, the same procedure was used to annotate the motifs using binding sequences from TRANSFAC and JASPAR. TRANSFAC contained binding sequences for 319 different TFs, which were mainly characterized in mammalian species. We chose a stringent threshold ($p < 10^{-5}$) and annotated 17,740 (11%) motifs as similar to 221 TRANSFAC TFBSs. The most commonly annotated TFBS was PAX5/BSAP: 969 motifs were similar to this site (Supplementary Table 3).

The annotation results using TFBSs from JASPAR overlapped substantially with the TRANSFAC results because the two databases use some of the same sources [33]. However, because the binding sequences in JASPAR were non-redundant, we chose a higher p-value threshold ($p < 10^{-4}$) for the JASPAR annotations, and annotated 23,331 (15%) motifs as similar to binding sites of 39 TFs. As with the TRANSFAC results, the most commonly annotated TFBS was BSAP/PAX5: 2041 motifs were similar to this site based on JASPAR binding sequence examples (Supplementary Table 3). In total, 40,396 (26%) motifs were annotated with at least one TFBS from one of the three databases.

3.2.6 cisRED Web Interface

All data and results discussed here, including orthologous upstream sequence region sets for each transcript, motifs found, and annotations, are available via the web interface at www.cisred.org [20]. URLs for motifs in figures are shown in Table 3.1. Additionally, all WABA and MotifSampler data are available on request. The features of the web interface are described below.

The main page allows the user to search the cisred motifs by exact transcript name (WormBase ID) or motif ID, search for analyzed transcripts in a specified genomic location, and search for motifs containing a given sequence (Figure 3.10).

The user can also browse transcript names in alphabetical order and annotated motifs by TF name. On the Gene View page the user can view information about an analyzed transcript followed by a list of motifs that were found upstream of that transcript.

This page shows the transcript and its upstream region in their genomic context via two embedded UCSC genome browser images. The list of motifs shows the genomic location, consensus sequence logo, and annotations (if applicable) for each motif, each of which are linked to the appropriate Motif View page.

Each motif has its own ID number and dedicated URL, which is linked back to the Gene View page belonging to the transcript the motif is associated with. On the Motif View page the user can see the genomic coordinates of the motif and its associated upstream region, a list of species the motif is found in, a list of annotations and their associated p-values (if applicable), the logo, the consensus sequence, the position frequency matrix, and the exact genomic coordinates and sequence for each species. On the Annotation page the user can view the motifs that were found to resemble a given TFBS.

The page shows the TFBS name and source (TRANSFAC, JASPAR, or ORegAnno) and the total number of motifs associated with this annotation. This is followed by a paged list of motifs, sorted by annotation p-value, each of which are linked to the appropriate Motif View and Gene View pages. In this way a user can quickly find motifs that strongly resemble a TFBS of interest.

3.2.7 Applications

Several examples of applications of the information in the cisRED *C. elegans* database to current questions in nematode genomics, gene annotation, evolution, and gene regulation are illustrated below.

Some wide motifs were unannotated protein-coding exons. There were 3918 motifs ≥ 30 bp wide. While many of these were in coding exons belonging to other transcripts of the same gene, others represented novel findings. Some of the wide motifs resembled protein-coding exons even though no coding exon was annotated by WormBase in that location. For example, a 120-bp motif was found immediately upstream of the ATG of *Y73B3A.12*, a member of the Calmodulin family (Figure 3.14 A). It had a depth of six species, occurring in all species except *C. briggsae* and *P. pacificus*. A BLASTX [34] search for the *C. elegans* motif sequence returned many matches to Calmodulin genes of various species, which indicated that this region of the *C. elegans* genome is likely to be a coding exon that was not annotated by WormBase.

Some highly conserved wide motifs may be noncoding RNA genes. A 143-bp motif was found upstream of *grd-7* (*F46H5.6*) (Figure 3.14 B), and all but five of the bases were perfectly conserved among four species of *Caenorhabditis* (this transcript had no acceptable *C. japonica* orthologue). The *C. briggsae* sequence included a one-bp insertion, causing a shift in the consensus sequence logo [31] at the 125th base of the motif. A BLAST search for this sequence returned no matches. However, WormBase indicated that the motif overlapped a predicted noncoding RNA gene near the 3' UTR of *unc-10* (*T10A3.1b*). This finding provides support for the predicted RNA gene in that location and its strong conservation in three other species suggests that it is functional. It also provides a hypothetical function for other very wide motifs that do not appear to be protein-coding.

Several very highly conserved motifs occurred in all eight nematode species. Thirteen transcripts had high-quality orthologues in all seven non-annotated species, and were associated with 115 motifs that occurred in all eight species. For example, a highly conserved motif was found in the 5' UTR of *rps-13* (*C16A3.9*) (Figure 3.14 C). Of the 12 bases that make up the motif, seven bases were perfectly conserved in all eight species.

Annotated motifs provided new information regarding TFBS locations and evolution of TF binding and function. The motif annotation process, which used TF binding sequences for both mammalian and *C. elegans* TFs, returned many novel binding site candidates. For example, a motif similar to a DAF-19 binding site was found near *kin-2* (*R07E4.6b*; Figure 3.14 D). The annotation results can also be used to suggest novel binding site candidates for uncharacterized TFs that are orthologues of characterized mammalian TFs. For example, a human ATF4-like motif was found near *Y34B4A.10* (Figure 3.14 E). Finally, the annotation process revealed information concerning the conservation of TFBSs in the more distant nematode species. For example, a DAF-19-like site near the uncharacterized gene *C54C6.6* (Figure 3.14 F) showed that the site was strongly conserved in *P. pacificus* and weakly conserved in *B. malayi*.

3.3 Discussion

The application of WABA to the seven non-*C. elegans* genomes revealed information about the recently sequenced genomes of *C. brenneri* and *C. japonica*. All four species in genus *Caenorhabditis* had similar overall numbers of matches to *C. elegans* WormPep sequences (Figure 3.3). However, compared to *C. briggsae* and *C. remanei*, there was a disproportionately small number of WormPep sequences that had one match and a large number with two matches in the *C. brenneri* genome. This anomaly may be because the draft genome sequence of *C. brenneri* is derived from a strain that was inbred and yet heterozygous over 30% of its genome. As alleles are highly differentiated in this species, the genome assembly contains alternative forms of many genes that were assembled independently [35]. *C. japonica* had 16% fewer matches to *C. elegans* WormPep protein-coding sequences than the other *Caenorhabditis* species, and had fewer high-quality orthologues. This may have been due to both the greater evolutionary distance between *C. elegans* and *C. japonica* and the poorer genome assembly of *C. japonica*, which was released very recently and was still in draft stages (Supplementary Table 1). High-quality orthologues among the more distant nematode species were even more rare; only 14% of examined *C. elegans* transcripts had high-quality orthologues in *Pristionchus pacificus*, *Brugia malayi*, or *Trichinella spiralis*. In addition to interference from the low quality of these genome assemblies, the WABA algorithm may be too stringent to find orthologues if the genomes are too distant. In order to minimize the impact of genomic anomalies and maximize the likelihood of finding evolutionarily conserved upstream motifs, we limited this investigation to transcripts with at least three high-quality orthologues. The resulting collection of orthologous upstream sequence region sets was strongly conserved and included only 17% of WormPep transcripts.

Of the 132 bidirectional promoters examined in this study, 86% were conserved among the species of genus *Caenorhabditis*. The majority of bidirectional promoters in *C. elegans* have previously been found to be conserved in *C. briggsae* [36]; given the high rate of conservation, bidirectional promoters must be an important mechanism for controlling gene regulation among gene pairs. Some gene pairs on bidirectional promoters are coexpressed while others have a mutually exclusive gene expression pattern [36]. Documentation of the conserved elements in these promoters, in combination with the examination of the expression patterns of the transcripts involved, may help to clarify these mechanisms of gene regulation.

While the large majority of orthologous regions in the other species were associated with only one *C. elegans* transcript, some functionally related groups of *C. elegans* transcripts had fewer orthologous representatives in the unannotated nematode genome sequences. Most cases of overlapping orthologues in the unannotated genomes belonged to large gene groups such as serpentine receptors. This may be because the four other species of *Caenorhabditis* are associated with different types of decaying matter [37]; *C. elegans* may have more of these types of receptors to help it find its specific type of food while the other species may have expanded different receptor families. In some cases, two *C. elegans* genes with overlapping orthologues were side by side (on the same or opposite strand), which suggests that a gene duplication event occurred in *C. elegans* after *C. elegans* diverged from the other *Caenorhabditis* species.

The transcripts that had a sufficient number of orthologues to be used in this analysis had a different chromosomal distribution from the entire set of WormPep transcripts, suggesting that certain regions of the genome are more highly conserved than others (Figure 3.5). Chromosome III is known to be rich in genes with yeast orthologues [9] and essential genes [38] such as those required for cell division [39]. A detailed analysis of synteny in the *C. elegans* and *C. briggsae* genomes has previously revealed that orthologues are overrepresented on Chromosomes III and X and underrepresented on Chromosome V [40].

Because regulatory elements are not readily distinguishable from other conserved upstream elements, the primary goal of this study was to catalogue all conserved elements of the upstream regions. We did not preface the motif discovery procedure with a multiple sequence alignment so as to avoid the preconditions that conserved elements be in the same order (with respect to the distance from the ATG) and contained within alignable sequence. We tested several motif discovery algorithms and found that while MotifSampler was the most suitable program for this purpose, a high-order background model was essential because nematode intergenic sequence frequently contains low-complexity sequence.

In order to assess the effectiveness of the motif discovery procedure, we compared discovered motifs to experimentally validated TFBSs from ORegAnno. The motif discovery algorithm was highly successful at finding experimentally validated sites, with a sensitivity of 82%. The upstream regions of the positive controls were only characterized with respect to locations of TFBSs (or predicted TFBSs; in some cases, the binding TF is not known). No sections of these upstream regions have been definitively shown not to have regulatory function. Because it is not possible to estimate the false positive rate without true negatives, we only used sensitivity and coverage to choose the threshold for motif inclusion.

We found 20 motifs upstream of *xbx-1* (*F02D8.3*), of which five were more highly conserved than the one corresponding to the DAF-19 site (Figure 3.9). Because functional analyses of promoter sequences tend to reveal only a few short TFBSs (see for example [4,6,27,32]), it seems unlikely that all of this conserved sequence has regulatory function. However, because the upstream sequence of *xbx-1* is uncharacterized other than the DAF-19 site, it is possible that some of the other motifs also have regulatory function.

While this study has focused on characterizing conserved elements, there is clearly much more to what constitutes a regulatory element than just conservation. Both TFBSs [41] and TFs [42] have been shown to be conserved among *C. elegans*, *C. briggsae*, and *C. remanei*. For the highly conserved transcripts studied here, we did not find regulatory elements to be more conserved than other portions of the upstream regions. There was no difference in the distribution of average IC between motifs overlapping experimentally validated sites and all motifs (Figure 3.8). Thus, attempts to assign a score to each motif indicating the probability that it had regulatory function were unsuccessful. In light of these results, we decided to retain all motifs that we identified, regardless of their conservation score.

Experimentally validated sites that were not found were poorly conserved or highly degenerate, and so were not reported by the motif discovery algorithm. Not all TFBSs were conserved; many of the experimentally validated sites had low IC while others were not found at all using our parameters for motif discovery. Additionally, some of the experimentally validated sites that our method did not identify may have been outside of the region we examined on the orthologous sequences, and there may be other ways to regulate transcription of the orthologues, perhaps using different TFs with a parallel function. The AT-rich sites such as PHA-4 [27] are highly degenerate and extremely common in the genome. Nematodes must have a way to distinguish functional from non-functional sites *in vivo*, perhaps via histone modifications [43].

In a preliminary comparison of conserved regions in *C. elegans* and *C. briggsae*, Siepel *et al.* [44] found that 18-37% of the genomes were conserved, but considered this to be an underestimate because they used phastCons-aligned regions. They anticipated that improved results could be generated by using additional nematode genomes. They suggested that highly conserved elements may contain multiple overlapping binding sites, be under protein-coding or RNA structural constraints, or have “as-yet-undiscovered functions”. They also suggested that some conserved regions may have “mutational rather than selectional explanations” and may be “shielded from mutations or subjected to hyperefficient repair”. The results described here were generated with eight nematode genomes. Consistent with their suggestion that alignment might underestimate conservation, we found that conserved elements identified using motif discovery resulted in a median coverage of 45% of the upstream regions. This proportion represents the amount of upstream sequence that was conserved to approximately the same degree as TFBSs, some of which are highly degenerate. Again consistent with their discussion, many wide motifs were in annotated or unannotated protein-coding exons belonging to the same gene. Protein-coding motifs can often be recognized by their codon-like conservation pattern in which every third base is poorly conserved because it can be substituted by several different nucleotides without changing the amino acid sequence (Figure 3.14 A); protein-coding regions also tend to have significant results following a BLASTX [34] search. Motifs that appear to be protein-coding but are not annotated could be used to refine *C. elegans* gene models. Some wide non-protein coding motifs were in 5' and 3' UTRs and may be target sites of RNA binding proteins or microRNAs, while others may represent noncoding RNA genes (Figure 3.14 B).

Most motifs were found in all sequences of the input set, with the result that most motifs have a species depth of four or five including *C. elegans*. The motif discovery algorithm preferred depth over conservation; if the best available version of the motif on one of the sequences was quite different from the others, it was included rather than excluded. This provided us with an opportunity to observe the evolution of conserved upstream elements among the more distant nematode species. Several motifs were found in all eight species and were very highly conserved (Figure 3.14 C), suggesting the presence of ancient genomic elements near essential genes.

Motifs for which the *C. elegans* sequence displayed a significant similarity to a characterized TFBS were annotated as such. We observed that conserved sequences similar to a wide variety of mammalian TFBSs appeared in *C. elegans* upstream regions. This annotation is preliminary and the intention was not to exhaustively annotate occurrences of TFBSs from TRANSFAC or JASPAR, but merely to assess which ones seemed to occur frequently among conserved parts of upstream regions as compared to non-conserved parts of upstream regions. There was substantial overlap between the annotation results using TRANSFAC and JASPAR, as JASPAR is a more thoroughly curated subset of TRANSFAC. The results from the two databases were consistent. For example, the most commonly annotated TF was the same for TRANSFAC and JASPAR (PAX5/BSAP) (Supplementary Table 3). Similarly, CREB was the fourth most commonly annotated TF from JASPAR and the third most commonly annotated TF from TRANSFAC.

Because certain characterized TFs in JASPAR, TRANSFAC, and ORegAnno had strongly variable or very few binding sequences, we chose to require a *C. elegans* sequence to be similar to a specific binding sequence rather than generate binding models such as position weight matrices for each TFBS. The limitation of this method was that all mismatches between the *C. elegans* sequence and a binding sequence were treated equally, which may have generated false positive annotations. Estimating the false positive rate requires a set of true negatives, and such a set is not available. Not all binding sites could be annotated using this method — some

TFs, such as PHA-4 and DAF-12, had so many variations in their binding sequences and were so common in the upstream regions that none of the motifs could be annotated with that TFBS at a p-value below the threshold. Motifs were much more likely than non-conserved upstream sequence to be similar to a TFBS. The distribution in scores between the motifs (by definition evolutionarily conserved) and non-conserved upstream sequence was different for most TFs.

A DAF-19-like site was found upstream of *kin-2* (*R07E4.6b*) (Figure 3.14 D). In addition to the high conservation of this site and its strong similarity to a DAF-19 binding site, we have further supporting evidence of its functionality. First, DAF-19 is known to regulate gene expression in ciliated neurons, and *kin-2* is expressed in ciliated neurons [45]. Secondly, KIN-2 is known to interact with RIC-8 [46], and *ric-8* (*Y69A2AR.2*) has been shown to be regulated by DAF-19 as well [41].

A human ATF4-like motif was found near *Y34B4A.10* (Figure 3.14 E). According to WormBase, the *C. elegans* homologue of the human *atf4* gene is *atf-5* (*T04C10.4*) [2]. The binding site of *C. elegans* ATF-5 is uncharacterized; perhaps conserved elements that are similar to the human ATF4 site could be tested for binding with, and regulation via, *C. elegans* ATF-5.

A DAF-19-like site was found upstream of the uncharacterized transcript *C54C6.6* (Figure 3.14 F). This site was shown to have substantial similarity in the more distant nematode species *P. pacificus* and *B. malayi*. The conservation of the site in these species suggests that they also have the DAF-19 TF and may use it to regulate the expression of some of the same genes. This example illustrates that annotated motifs can increase our understanding of gene regulation in these species.

3.3.1 Conclusions

We have shown that WABA is an effective tool for finding orthologues for highly conserved transcripts among nematode genomes. We applied WABA to all annotated protein-coding transcripts from *C. elegans*; however, only transcripts with at least three high-quality orthologues were included in the motif discovery step. We identified conserved elements in the upstream regions of 3847 *C. elegans* transcripts (17% of all *C. elegans* transcripts).

We found that identification of putative regulatory elements via motif discovery among orthologous upstream regions resulted in a sensitivity of 82%, which suggests that most regulatory elements are conserved. However, we also found that the upstream regions also contain numerous other conserved elements, and that regulatory elements are not the most highly conserved elements in these upstream regions. Therefore, while conserved motifs are enriched for regulatory elements, conservation alone can not be used to distinguish regulatory elements from other conserved elements.

All of our results are publicly available via the web interface at www.cisred.org. Gene regulation researchers can use the web interface to see all conserved elements and their annotations for any gene of interest. For work involving laboratory methods such as reporter gene assays and gel shift assays to investigate the regulation of these genes, the cisRED data can immediately focus the search onto conserved and possibly annotated elements in upstream regions.

Many of the conserved elements in the cisRED database are in 5' and 3' UTRs of different transcripts; some of these may be candidate targets for RNA binding proteins. Additionally, some of the wide, highly conserved motifs may serve as novel noncoding RNA gene candidates. Those motifs that appear to be protein-coding can be used to refine and expand existing gene models.

Twenty-six percent of the conserved elements were found to be similar to known TFBSs and were annotated as such. These annotations are useful in three important ways. First, they provide novel candidate binding sites for TFs that are already characterized in *C. elegans*. These sites could be tested by researchers who are interested in targets of the TFs. Secondly, the annotations provide novel binding site candidates for uncharacterized TFs that are orthologues of characterized mammalian TFs. This takes advantage of existing information about TF binding in mammals to expand our understanding of gene regulation in *C. elegans*. Lastly, the annotations make it possible to assess evolution of TFs, their binding sites, and the process of gene regulation in general by comparing both the TF protein sequence and their predicted binding sites across the different nematode species. The conservation of annotated sites in more distantly related nematodes implies that they have the same TFs as *C. elegans* and use them to regulate some of the same genes.

3.4 Methods

3.4.1 Orthologue Identification

Genome sequences were obtained from the WormBase and Washington University FTP servers (Supplementary Table 1). WS170 was used because the cisRED web interface makes extensive use of the UCSC Genome Browser and that was the version of the *C. elegans* genome at UCSC as of May 2008. WABA [18] was used to find one or more orthologous sequences in each of the other genomes for each of the 23,212 chromosomal protein-coding transcripts in

WormPep. Only single alignments from WABA that aligned beginning at the ATG of the *C. elegans* sequence (i.e. 'high-quality orthologues') were retained.

3.4.2 Orthologous Upstream Sequence Regions

The upstream region of each *C. elegans* WormPep transcript was combined with the upstream regions of its orthologues in the other nematode genomes to form an orthologous upstream sequence region set. Only transcripts that had at least three out of a possible seven high-quality orthologues were used. Of the 192 curated *C. elegans* TFBSs in ORegAnno, 83% were within 1500 bp of the ATG. The remaining TFBSs were sparsely distributed up to 9 kbp upstream and up to 9 kbp downstream of the ATG; the region further upstream than 1500 bp was not enriched for TFBSs. Half of *C. elegans* transcripts had another gene within 1500 bp of the ATG. The upstream sequence used was defined as 1500 bp upstream of the ATG (including the 5' UTR, if present) or up to the end of the nearest protein-coding transcript, WABA match, or end of contig. The 1500 bp excluded masked repeats and undefined sequence (Ns), and was limited to a maximum total length of 3000 bp. A minimum of 100 bp was required for *C. elegans* to avoid transcripts whose upstream region was too short to analyze efficiently. We excluded 59 *C. elegans* transcripts for this reason; of these the closest upstream transcript was on the same strand for 28 and on the opposite strand for 31.

3.4.3 Motif Discovery

We applied the motif discovery algorithms MEME [47], CONSENSUS [29], and MotifSampler [19] to the upstream sets and compared their relative performance in detecting a set of experimentally discovered TFBSs obtained from ORegAnno. Of the three methods, only MotifSampler could detect the positive controls with greater than 25% sensitivity and combining the results of two or more methods did not improve the sensitivity. Consequently, we used only MotifSampler to detect motifs in the orthologous upstream sets. For each orthologous upstream sequence region set, a background sequence set was generated that contained randomly selected upstream sequences from each species in the same proportions as the foreground sequences. A third-order Markov background model was then generated from each background sequence set.

MotifSampler was run using the following parameters: -p 0.3 -s 1 -n 25 -r 30. The 'r' parameter specifies 30 iterations on each sequence set; we used the score assigned to each motif by MotifSampler to retain the top 30% of motifs from each sequence set. Motif discovery was performed using target widths of 6, 8, 10, 12, and 14 bp because 86% of *C. elegans* TFBSs in ORegAnno are in this width range. Motifs that overlapped consistently on all sequences on which they were found were merged into one motif. Motifs for which MotifSampler returned multiple instances on the *C. elegans* sequence were separated and matched with the most conserved instance of that motif on each orthologous sequence. Motifs that occurred on the orthologous sequences but not on the *C. elegans* sequence were discarded. Each motif in the cisRED database is an aligned collection of sequences containing one sequence from the *C. elegans* upstream region and not more than one sequence from each orthologous upstream region.

3.4.4 Validation

Experimentally validated TFBSs from ORegAnno [21] were used as positive controls for motif discovery. ORegAnno contains 192 TFBSs for *C. elegans*, of which 44 were found in 28 of the upstream regions of this analysis. An experimentally validated TFBS from ORegAnno was considered to be discovered when the predicted motif overlapped at least 50% of the site. The average information content (IC) of each motif was calculated as described by Hertz and Stormo [29].

3.4.5 Annotation to Show Similarity to Known TFBSs

Binding sequences for characterized TFs were obtained from TRANSFAC (version 9.2) [22], JASPAR (version 4) [23], and ORegAnno. Each TF in these databases was associated with a set of between one and 179 sequences that had been experimentally shown to bind that TF.

The *C. elegans* sequence of each motif was compared with each database TF and scored as follows. The score between the *C. elegans* sequence and a single binding sequence was the number of mismatches between the two sequences divided by the width of the binding sequence. We required a minimum overlap of five bp between the motif and the binding sequence; flanking genomic sequence was included as needed. We retained the minimum score with respect to relative strand orientation and position of the two sequences, and the minimum such score over all of the TF's binding sequences.

We assigned a p-value to the retained score for each motif-TF pair based on the background score distribution of that TF, which we generated by scoring 1000 randomly chosen *C. elegans* upstream sequences that were not covered by motifs. Motifs were annotated as similar to a binding site if the p-value of the motif-TF score was below a threshold as follows: ORegAnno binding sites: p-value threshold = 0.00015; TRANSFAC binding sites: p-value threshold = 0.00001; JASPAR binding sites: p-value threshold = 0.0001.

3.5 Chapter 3 Table

Figure	URL
3.7 A	http://www.cisred.org/c.elegans4/siteseq?fid=157071
3.7 B	http://www.cisred.org/c.elegans4/siteseq?fid=130462
3.7 C	http://www.cisred.org/c.elegans4/siteseq?fid=92832
3.9	http://www.cisred.org/c.elegans4/gene_view?ensembl_id=F02D8.3
3.10	http://www.cisred.org/c.elegans4/
3.11	http://www.cisred.org/c.elegans4/gene_view?ensembl_id=T27B1.1
3.12	http://www.cisred.org/c.elegans4/siteseq?fid=126133
3.13	http://www.cisred.org/c.elegans4/c.elegans4/group_content_view?aid=30009
3.14 A	http://www.cisred.org/c.elegans4/siteseq?fid=151292
3.14 B	http://www.cisred.org/c.elegans4/siteseq?fid=71907
3.14 C	http://www.cisred.org/c.elegans4/siteseq?fid=17781
3.14 D	http://www.cisred.org/c.elegans4/siteseq?fid=102892
3.14 E	http://www.cisred.org/c.elegans4/siteseq?fid=136618
3.14 F	http://www.cisred.org/c.elegans4/siteseq?fid=37257

Table 3.1 – Figure URLs

All results are available via the cisRED web interface. URLs of motifs in figures are indicated.

3.6 Chapter 3 Figures

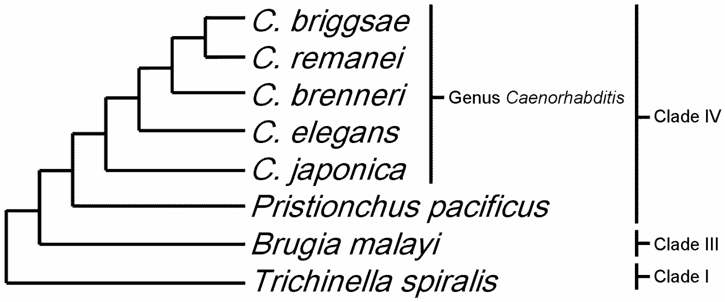


Figure 3.1 – Phylogenetic Tree of Species

C. briggsae, *C. remanei*, and *C. brenneri* are all more closely related to each other than they are to *C. elegans*, while C. japonica is an outgroup within genus *Caenorhabditis*. *Pristionchus pacificus*, like *C. elegans*, is a hermaphroditic bacteriovore and belongs to the same clade of nematodes as *C. elegans*, but *Brugia malayi* and *Trichinella spiralis* are mammalian parasites from other clades in phylum Nematoda. Evolutionary distances are not to scale.

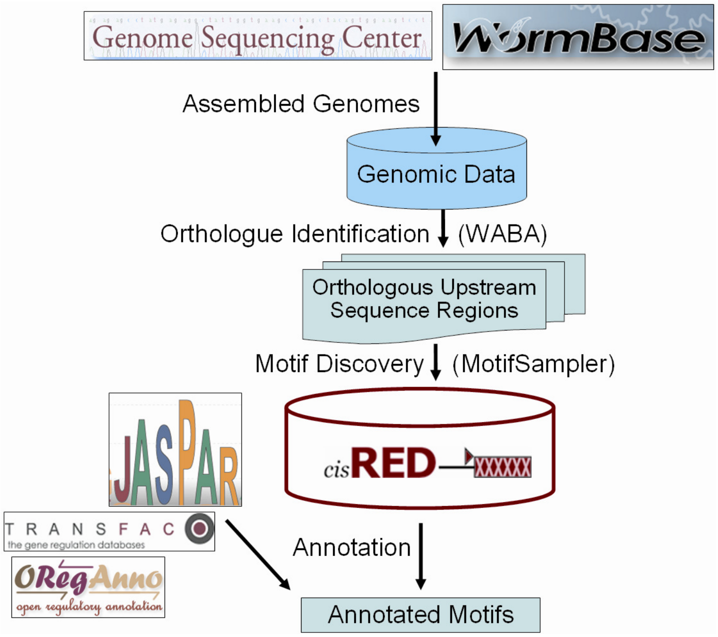


Figure 3.2 – Flowchart of Method

Assembled genomes for the eight genomes and WormPep sequences for *C. elegans* were obtained from FTP servers at WormBase and Washington University Genome Sequence Center. Orthologues of *C. elegans* protein-coding transcripts in the other seven genomes were inferred using WABA. The upstream region of each *C. elegans* transcript was pooled with the upstream regions of its orthologues from the other genomes to form orthologous upstream sequence region sets. Only *C. elegans* transcripts with at least three high-quality orthologues were used for this analysis. The motif discovery algorithm MotifSampler was used to find conserved elements in each orthologous upstream sequence region set, which were placed in the cisRED *C. elegans* database. Motifs were examined for similarity to experimentally validated transcription factor binding sequences from TRANSFAC, JASPAR, and ORegAnno, and those motifs with a significant resemblance to a binding sequence were annotated as such. All motifs and their annotations are publicly available via the web interface at www.cisred.org.

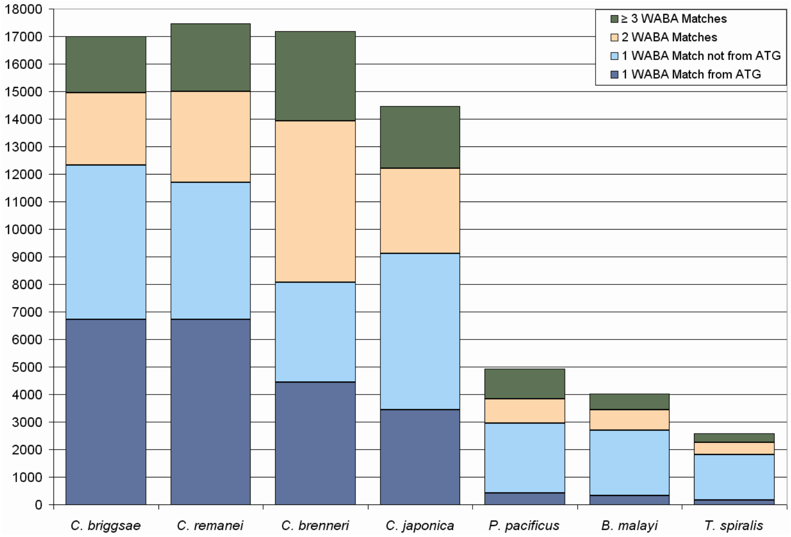


Figure 3.3 – Number of WABA Matches for 23,212 Chromosomal *C. elegans* WormPep Transcripts

The number of C. elegans transcripts with exactly one match starting from the ATG ('high-quality orthologues') is shown at the bottom, in dark blue. The number of remaining C. elegans transcripts with exactly one match is shown in light blue. The number of C. elegans transcripts with two matches in the comparison genome is shown in yellow, and the number of C. elegans transcripts with three or more matches is shown in green.

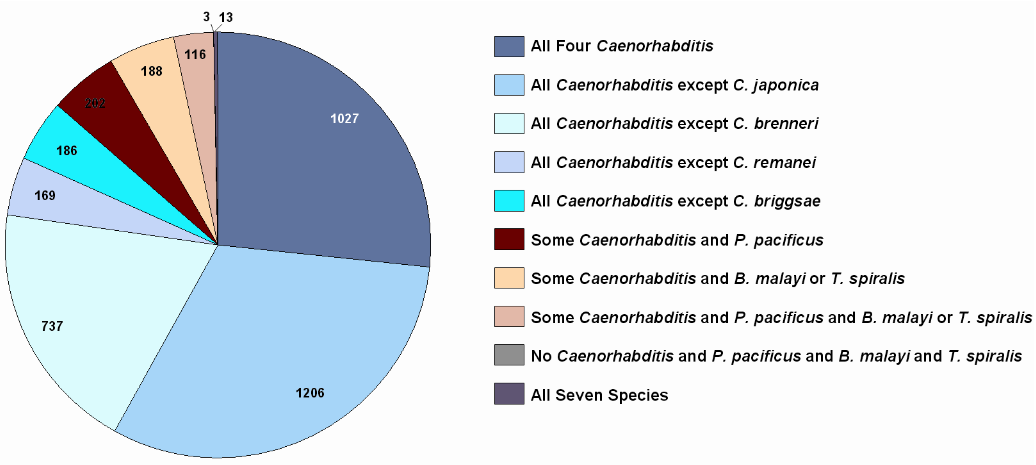


Figure 3.4 – Species Composition of Orthologous Upstream Sequence Region Sets

The upstream regions of *C. elegans* transcripts were pooled with the upstream regions of their orthologues from the other seven genomes to form orthologous upstream sequence region sets. Only *C. elegans* transcripts with at least three high-quality orthologues were used, resulting in a total of 3847 sets. Of these, 1027 contained sequence from all four species in genus *Caenorhabditis* (dark blue), while a total of 2298 of the sets contained sequence from all but one of the four *Caenorhabditis* species (various shades of light blue). Only 522 of the sets contained sequence from *Pristionchus pacificus*, *Brugia malayi*, or *Trichinella spiralis*; 13 sets contained sequence from all seven species (purple).

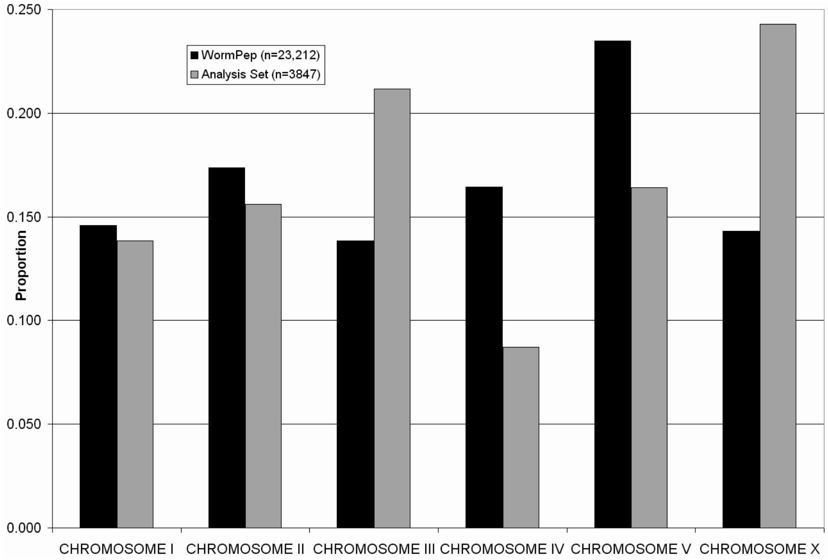


Figure 3.5 – Chromosomal Distribution of Orthologous Upstream Sequence Region Sets

The proportion of WormPep transcripts on each of the six chromosomes (black), and the proportion of transcripts used in this analysis (grey).

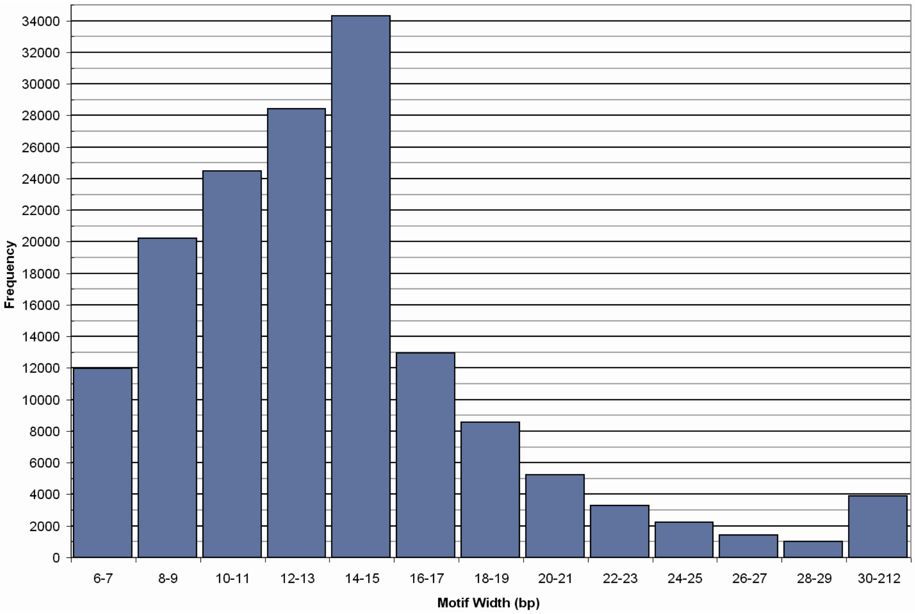


Figure 3.6 – Histogram of Motif Widths

MotifSampler was used to find motifs of width 6, 8, 10, 12, and 14 bp. Motifs that overlapped in all sequences were merged to form wider motifs, with the result that 3918 motifs were ≥ 30 bp wide and the widest motif was 212 bp.

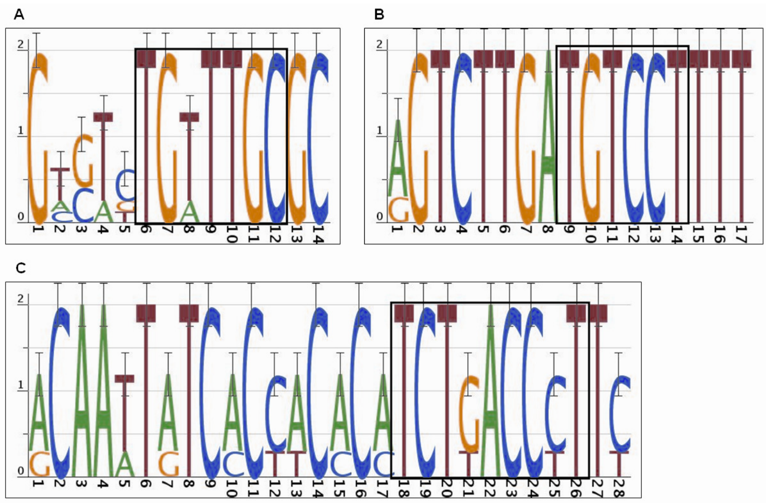


Figure 3.7 – Examples of Experimentally Validated Sites

A – A motif that overlaps a PHA-4 site upstream of *tph-1* (*ZK1290.2b*); B – A motif that overlaps a DAF-12 site upstream of *lit-1* (*W06F12.1c*); C – A motif that overlaps an ‘Early-2’ site upstream of *K07C11.4*. Locations of experimentally validated sites are indicated by black boxes. cisRED URLs are indicated in Table 3.1.

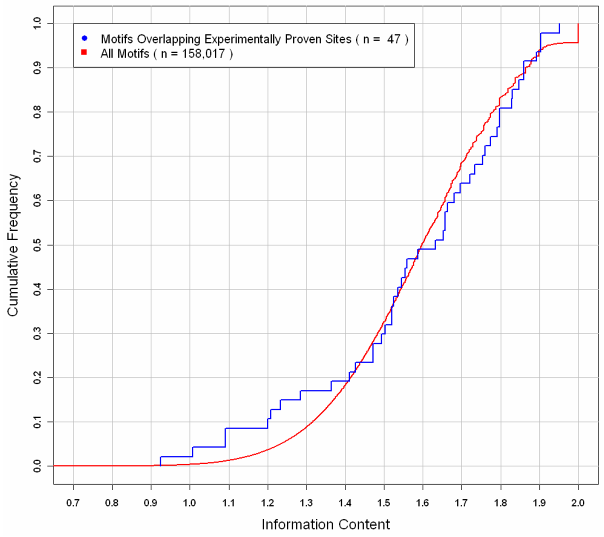


Figure 3.8 – Cumulative Distribution of Information Content of Motifs

The cumulative distribution functions of the average information content for all motifs (red) and for motifs that overlapped experimentally validated TFBSs from ORegAnno (blue).

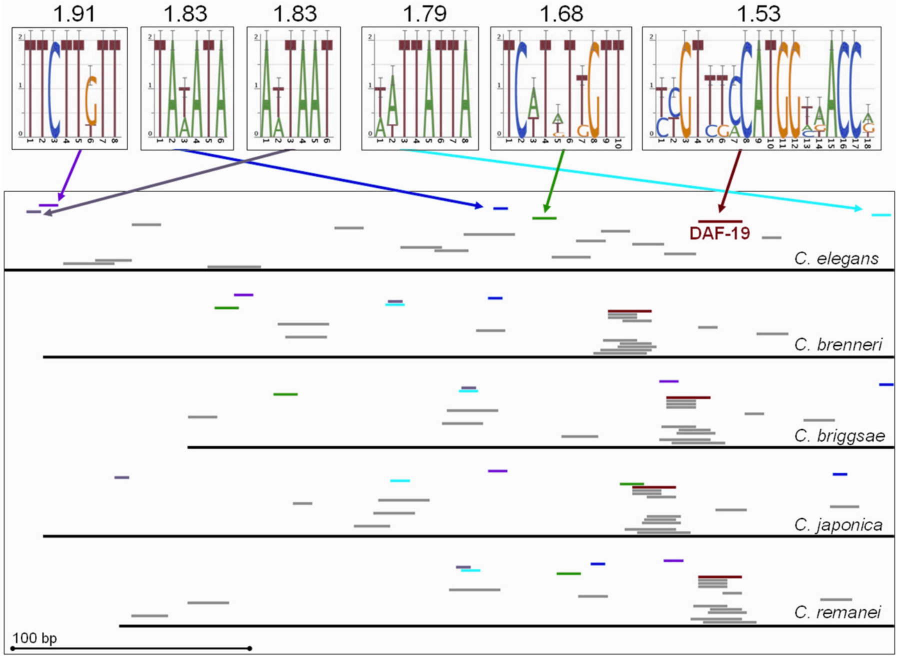
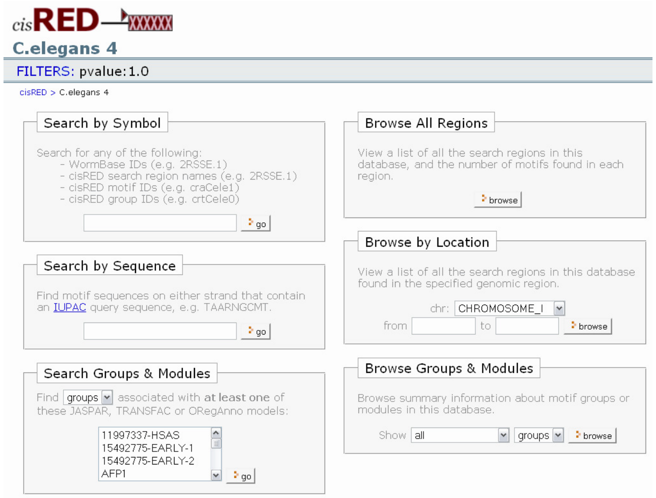


Figure 3.9 – Example of High-coverage Upstream Sequence Region with an Experimentally Validated Site

The upstream regions of *xbx-1* (*F02D8.3*) and its orthologues in *C. briggsae*, *C. remanei*, *C. brenneri*, and *C. japonica* are indicated by black lines. The ATG of each transcript or putative orthologue is at the right edge of the figure. The logos of the top six most-conserved motifs and their IC are shown; the locations of these motifs in each upstream sequence are indicated by coloured bars. The locations of the remaining motifs are indicated by grey bars. Motifs are sorted by IC with the most conserved motif at the top. The experimentally validated DAF-19 site is indicated. The cisRED URL is indicated in Table 3.1.



Search for any of the following:
- WormBase IDs (e.g. 2RSSE.1)
- cisRED search region names (e.g. 2RSSE.1)
- cisRED motif IDs (e.g. craCele1)
- cisRED group IDs (e.g. crtCele0)

Find motif sequences on either strand that contain an IUPAC query sequence, e.g. TAARNGCMT.

Find groups associated with **at least one** of these JASPAR, TRANSFAC or ORegAnno models:

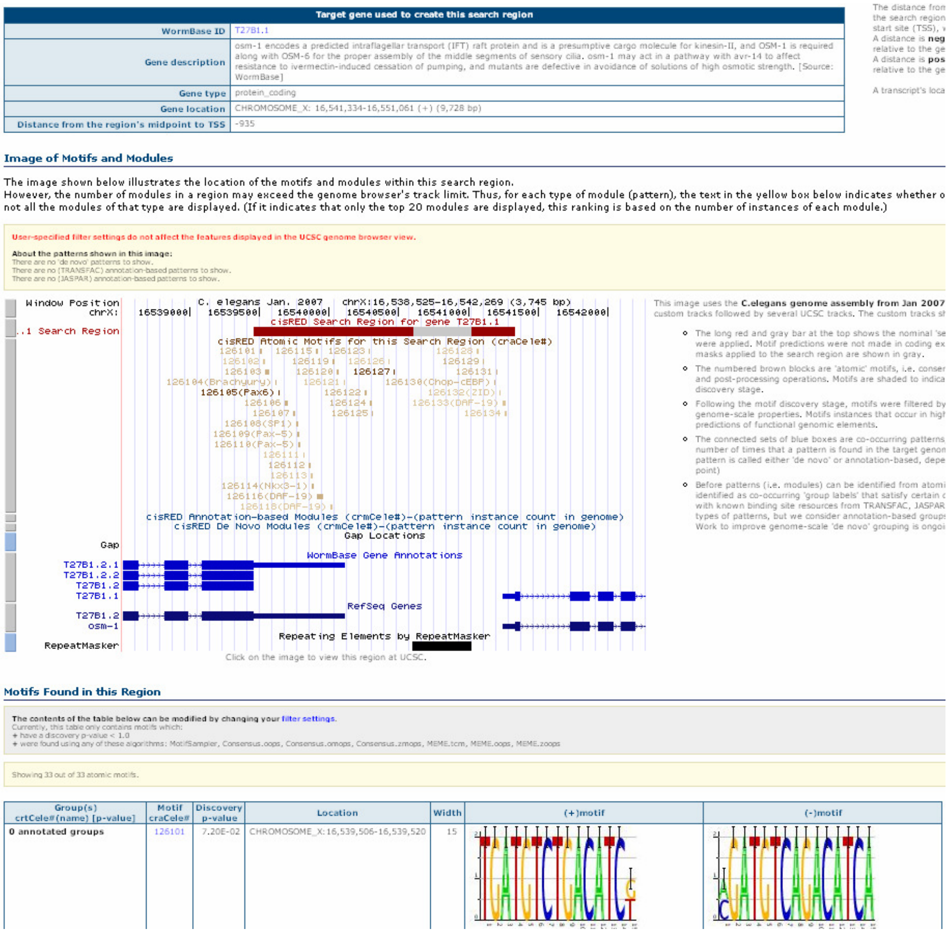
11997337-HSAS
15492775-EARLY-1
15492775-EARLY-2
AFP1

View a list of all the search regions in this database, and the number of motifs found in each region.

View a list of all the search regions in this database found in the specified genomic region.

Browse summary information about motif groups or modules in this database.

Figure 3.10 – The cisRED *C. elegans* Web Interface Main Page


Window Position
chrX:
C. elegans Jan. 2007   chrX:16,538,525-16,542,269 (3,745 bp)
16539000| 16539500| 16540000| 16540500| 16541000| 16541500| 16542000|
..1 Search Region
cisRED Search Region for gene T27B1.1
cisRED Atomic Motifs for this Search Region (craCele#)
126101 126115 126123 126128
126102 126119 126126 126129
126103 126120 126127 126131
126104(Brachyury) 126121 126130(Chop-cEBP)
126105(Pax6) 126122 126132(ZID)
126106 126124 126133(DAF-19)
126107 126125 126134
126108(SP1)
126109(Pax-5)
126110(Pax-5)
126111
126112
126113
126114(Nkx3-1)
126116(DAF-19)
126118(DAF-19)
cisRED Annotation-based Modules (crmCele#)-(pattern instance count in genome)
cisRED De Novo Modules (crmCele#)-(pattern instance count in genome)
Gap Locations
Gap
WormBase Gene Annotations
T27B1.2.1
T27B1.2.2
T27B1.2
T27B1.1
RefSeq Genes
T27B1.2
osm-1
Repeating Elements by RepeatMasker
RepeatMasker
2
1
0

Target gene used to create this search region	
WormBase ID	T27B1.1
Gene description	osm-1 encodes a predicted intraflagellar transport (IFT) raft protein and is a presumptive cargo molecule for kinesin-II, and OSM-1 is required along with OSM-6 for the proper assembly of the middle segments of sensory cilia. osm-1 may act in a pathway with avr-14 to affect resistance to ivermectin-induced cessation of pumping, and mutants are defective in avoidance of solutions of high osmotic strength. [Source: WormBase]
Gene type	protein_coding
Gene location	CHROMOSOME_X: 16,541,334-16,551,061 (+) (9,728 bp)
Distance from the region's midpoint to TSS	-935

The distance fron
the search region
start site (TSS), v
A distance is **neg**
relative to the ge
A distance is **pos**
relative to the ge

A transcript's loca

Image of Motifs and Modules

The image shown below illustrates the location of the motifs and modules within this search region.
However, the number of modules in a region may exceed the genome browser's track limit. Thus, for each type of module (pattern), the text in the yellow box below indicates whether o
not all the modules of that type are displayed. (If it indicates that only the top 20 modules are displayed, this ranking is based on the number of instances of each module.)

User-specified filter settings do not affect the features displayed in the UCSC genome browser view.

About the patterns shown in this image:
There are no 'de novo' patterns to show.
There are no (TRANSFAC) annotation-based patterns to show.
There are no (JASPAR) annotation-based patterns to show.

This image uses the **C.elegans genome assembly from Jan 2007**
custom tracks followed by several UCSC tracks. The custom tracks sh

- The long red and gray bar at the top shows the nominal 'se
were applied. Motif predictions were not made in coding ex
masks applied to the search region are shown in gray.
- The numbered brown blocks are 'atomic' motifs, i.e. conser
and post-processing operations. Motifs are shaded to indica
discovery stage.
- Following the motif discovery stage, motifs were filtered by
genome-scale properties. Motifs instances that occur in high
predictions of functional genomic elements.
- The connected sets of blue boxes are co-occurring patterns
number of times that a pattern is found in the target genon
pattern is called either 'de novo' or annotation-based, depe
point)
- Before patterns (i.e. modules) can be identified from atomi
identified as co-occurring 'group labels' that satisfy certain c
with known binding site resources from TRANSFAC, JASPAR
types of patterns, but we consider annotation-based groups
Work to improve genome-scale 'de novo' grouping is ongoi

Click on the image to view this region at UCSC.

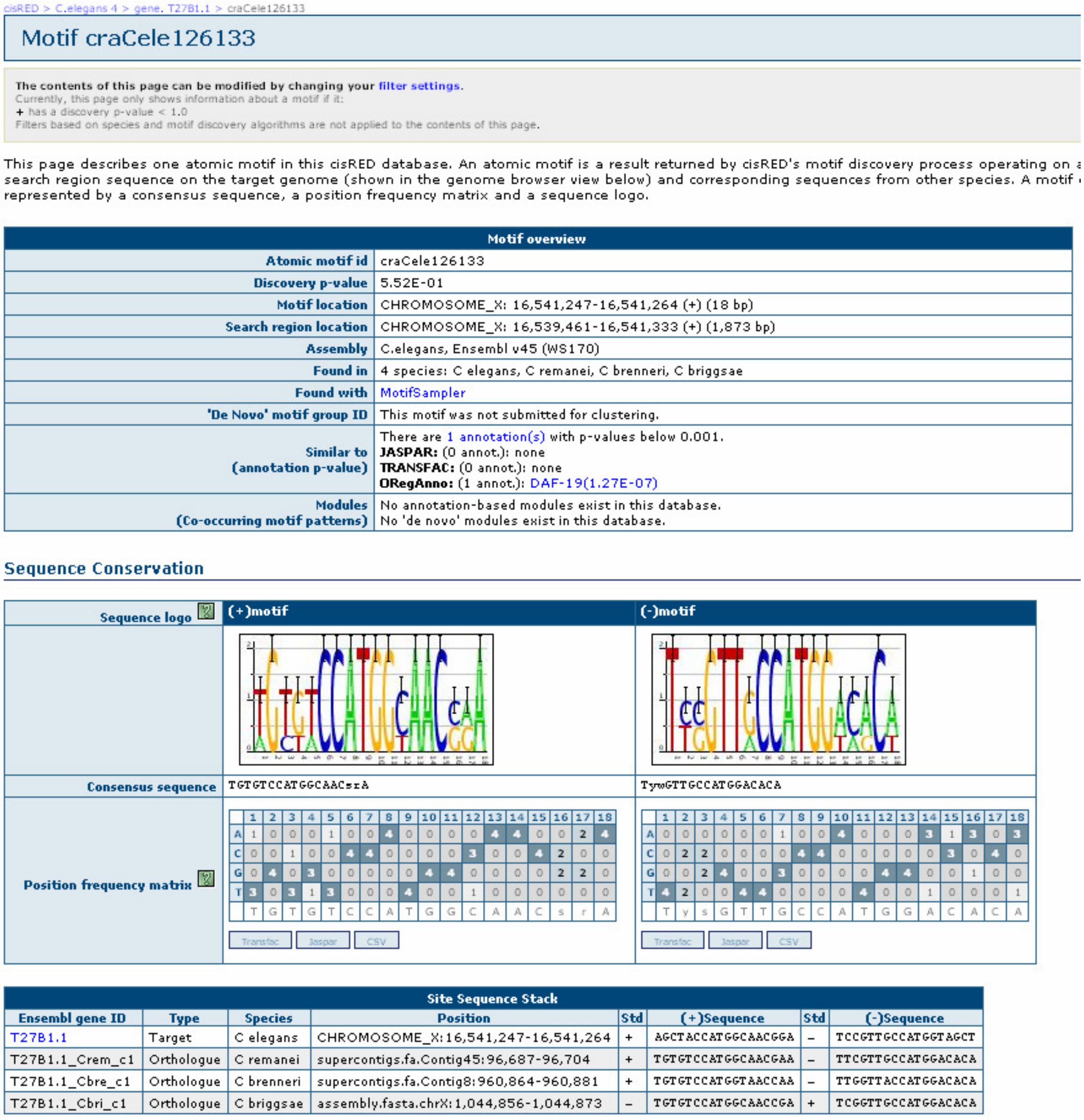
Motifs Found in this Region

The contents of the table below can be modified by changing your filter settings.
Currently, this table only contains motifs which:
+ have a discovery p-value < 1.0
+ were found using any of these algorithms: MotifSampler, Consensus.oops, Consensus.omops, Consensus.zmops, MEME.tcm, MEME.oops, MEME.zoops

Showing 33 out of 33 atomic motifs.

Group(s) crtCele#(name) [p-value]	Motif craCele#	Discovery p-value	Location	Width	(+)motif	(-)motif
0 annotated groups	126101	7.20E-02	CHROMOSOME_X:16,539,506-16,539,520	15		

Figure 3.11 – An Example of a cisRED *C. elegans* Web Interface Gene View Page.



cisRED > C.elegans 4 > gene. T27B1.1 > craCele126133

Motif craCele126133

The contents of this page can be modified by changing your filter settings.
Currently, this page only shows information about a motif if it:
+ has a discovery p-value < 1.0
Filters based on species and motif discovery algorithms are not applied to the contents of this page.

This page describes one atomic motif in this cisRED database. An atomic motif is a result returned by cisRED's motif discovery process operating on a search region sequence on the target genome (shown in the genome browser view below) and corresponding sequences from other species. A motif represented by a consensus sequence, a position frequency matrix and a sequence logo.

Motif overview	
Atomic motif id	craCele126133
Discovery p-value	5.52E-01
Motif location	CHROMOSOME_X: 16,541,247-16,541,264 (+) (18 bp)
Search region location	CHROMOSOME_X: 16,539,461-16,541,333 (+) (1,873 bp)
Assembly	C.elegans, Ensembl v45 (WS170)
Found in	4 species: C elegans, C remanei, C brenneri, C briggsae
Found with	MotifSampler
'De Novo' motif group ID	This motif was not submitted for clustering.
Similar to (annotation p-value)	There are 1 annotation(s) with p-values below 0.001. **JASPAR:** (0 annot.): none **TRANSFAC:** (0 annot.): none **ORegAnno:** (1 annot.): DAF-19(1.27E-07)
Modules (Co-occuring motif patterns)	No annotation-based modules exist in this database. No 'de novo' modules exist in this database.

Sequence Conservation

Sequence logo ? (+)motif (-)motif

Consensus sequence

(+)motif: TGTGTCCATGGCAACsrA

(-)motif: TywGTTGCCATGGACACA

Position frequency matrix ?

(+)motif:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	1	0	0	0	1	0	0	4	0	0	0	0	4	4	0	0	2	4
C	0	0	1	0	0	4	4	0	0	0	0	3	0	0	4	2	0	0
G	0	4	0	3	0	0	0	0	0	4	4	0	0	0	0	2	2	0
T	3	0	3	1	3	0	0	0	4	0	0	1	0	0	0	0	0	0
	T	G	T	G	T	C	C	A	T	G	G	C	A	A	C	s	r	A

Transfac Jaspar CSV

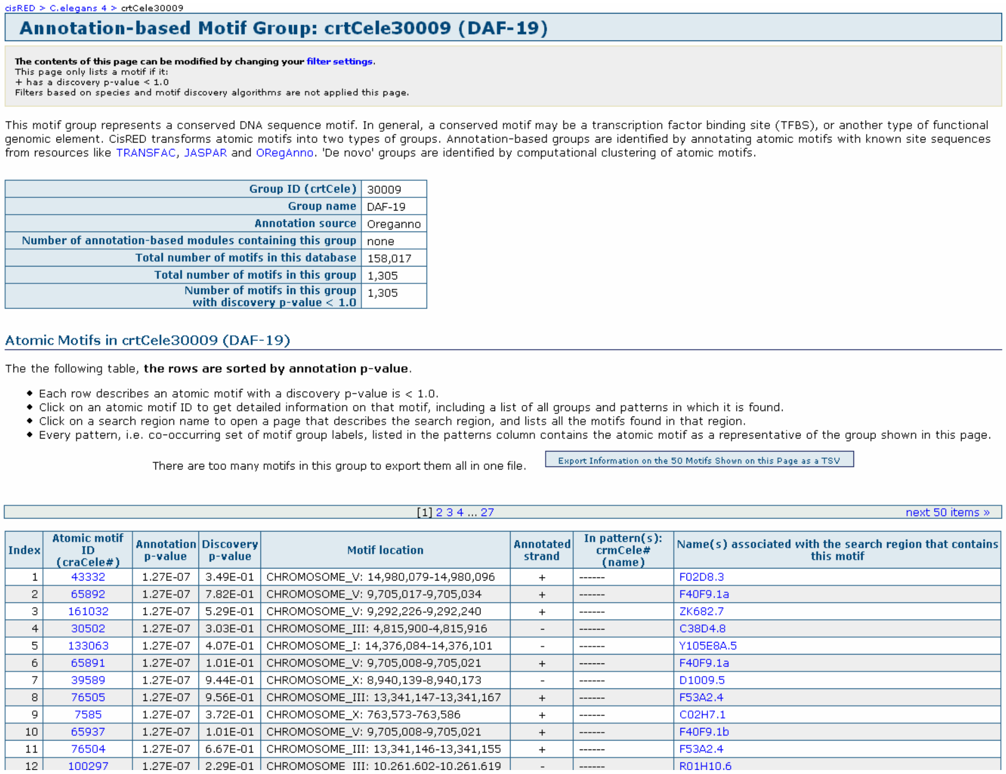
(-)motif:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	0	0	0	0	0	0	1	0	0	4	0	0	0	3	1	3	0	3
C	0	2	2	0	0	0	0	4	4	0	0	0	0	0	3	0	4	0
G	0	0	2	4	0	0	3	0	0	0	0	4	4	0	0	1	0	0
T	4	2	0	0	4	4	0	0	0	0	4	0	0	1	0	0	0	1
	T	y	s	G	T	T	G	C	C	A	T	G	G	A	C	A	C	A

Transfac Jaspar CSV

Site Sequence Stack							
Ensembl gene ID	**Type**	**Species**	**Position**	**Std**	**(+)Sequence**	**Std**	**(-)Sequence**
T27B1.1	Target	C elegans	CHROMOSOME_X:16,541,247-16,541,264	+	AGCTACCATGGCAACGGA	–	TCCGTTGCCATGGTAGCT
T27B1.1_Crem_c1	Orthologue	C remanei	supercontigs.fa.Contig45:96,687-96,704	+	TGTGTCCATGGCAACGAA	–	TTCGTTGCCATGGACACA
T27B1.1_Cbre_c1	Orthologue	C brenneri	supercontigs.fa.Contig8:960,864-960,881	+	TGTGTCCATGGTAACCAA	–	TTGGTTACCATGGACACA
T27B1.1_Cbri_c1	Orthologue	C briggsae	assembly.fasta.chrX:1,044,856-1,044,873	–	TGTGTCCATGGCAACCGA	+	TCGGTTGCCATGGACACA

Figure 3.12 – An Example of a cisRED *C. elegans* Web Interface Motif View Page.



Annotation-based Motif Group: crtCele30009 (DAF-19)

The contents of this page can be modified by changing your filter settings.
This page only lists a motif if it:
+ has a discovery p-value < 1.0
Filters based on species and motif discovery algorithms are not applied this page.

This motif group represents a conserved DNA sequence motif. In general, a conserved motif may be a transcription factor binding site (TFBS), or another type of functional genomic element. CisRED transforms atomic motifs into two types of groups. Annotation-based groups are identified by annotating atomic motifs with known site sequences from resources like TRANSFAC, JASPAR and ORegAnno. 'De novo' groups are identified by computational clustering of atomic motifs.

Group ID (crtCele)	30009
Group name	DAF-19
Annotation source	Oreganno
Number of annotation-based modules containing this group	none
Total number of motifs in this database	158,017
Total number of motifs in this group	1,305
Number of motifs in this group with discovery p-value < 1.0	1,305

Atomic Motifs in crtCele30009 (DAF-19)

The the following table, **the rows are sorted by annotation p-value**.

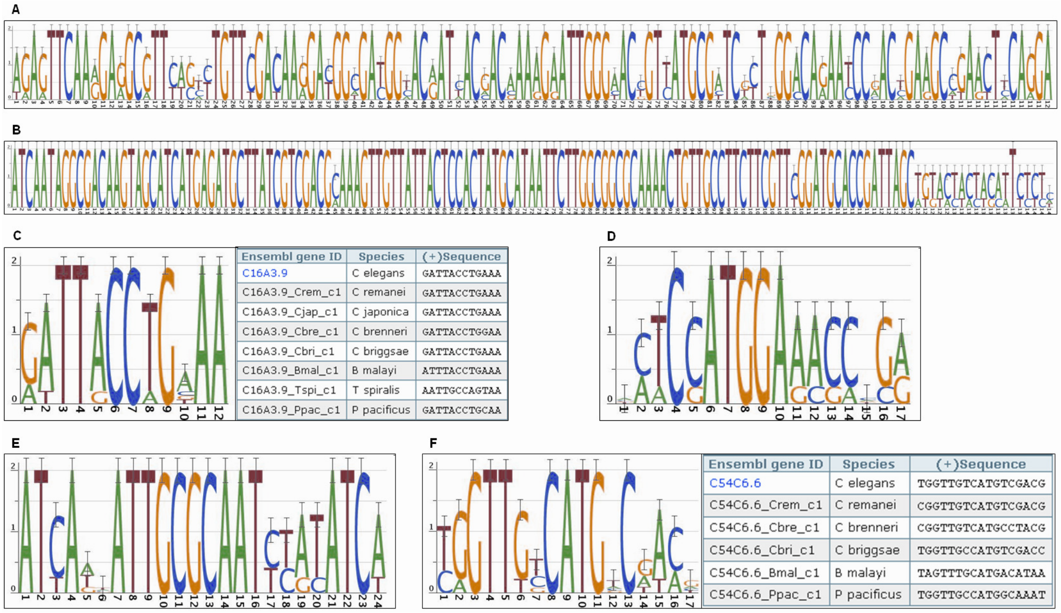
- Each row describes an atomic motif with a discovery p-value is < 1.0.
- Click on an atomic motif ID to get detailed information on that motif, including a list of all groups and patterns in which it is found.
- Click on a search region name to open a page that describes the search region, and lists all the motifs found in that region.
- Every pattern, i.e. co-occurring set of motif group labels, listed in the patterns column contains the atomic motif as a representative of the group shown in this page.

There are too many motifs in this group to export them all in one file.

Export Information on the 50 Motifs Shown on this Page as a TSV
[1] 2 3 4 ... 27
next 50 items »

Index	Atomic motif ID (craCele#)	Annotation p-value	Discovery p-value	Motif location	Annotated strand	In pattern(s): crmCele# (name)	Name(s) associated with the search region that contains this motif
1	43332	1.27E-07	3.49E-01	CHROMOSOME_V: 14,980,079-14,980,096	+	------	F02D8.3
2	65892	1.27E-07	7.82E-01	CHROMOSOME_V: 9,705,017-9,705,034	+	------	F40F9.1a
3	161032	1.27E-07	5.29E-01	CHROMOSOME_V: 9,292,226-9,292,240	+	------	ZK682.7
4	30502	1.27E-07	3.03E-01	CHROMOSOME_III: 4,815,900-4,815,916	-	------	C38D4.8
5	133063	1.27E-07	4.07E-01	CHROMOSOME_I: 14,376,084-14,376,101	-	------	Y105E8A.5
6	65891	1.27E-07	1.01E-01	CHROMOSOME_V: 9,705,008-9,705,021	+	------	F40F9.1a
7	39589	1.27E-07	9.44E-01	CHROMOSOME_X: 8,940,139-8,940,173	-	------	D1009.5
8	76505	1.27E-07	9.56E-01	CHROMOSOME_III: 13,341,147-13,341,167	+	------	F53A2.4
9	7585	1.27E-07	3.72E-01	CHROMOSOME_X: 763,573-763,586	+	------	C02H7.1
10	65937	1.27E-07	1.01E-01	CHROMOSOME_V: 9,705,008-9,705,021	+	------	F40F9.1b
11	76504	1.27E-07	6.67E-01	CHROMOSOME_III: 13,341,146-13,341,155	+	------	F53A2.4
12	100297	1.27E-07	2.29E-01	CHROMOSOME_III: 10,261,602-10,261,619	-	------	R01H10.6

Figure 3.13 – An Example of a cisRED *C. elegans* Web Interface Annotation View Page.



Ensembl gene ID	Species	(+)Sequence
C16A3.9	C elegans	GATTACCTGAAA
C16A3.9_Crem_c1	C remanei	GATTACCTGAAA
C16A3.9_Cjap_c1	C japonica	GATTACCTGAAA
C16A3.9_Cbre_c1	C brenneri	GATTACCTGGAA
C16A3.9_Cbri_c1	C briggsae	GATTACCTGAAA
C16A3.9_Bmal_c1	B malayi	ATTTACCTGAAA
C16A3.9_Tspi_c1	T spiralis	AATTGCCAGTAA
C16A3.9_Ppac_c1	P pacificus	GATTACCTGCAA

Ensembl gene ID	Species	(+)Sequence
C54C6.6	C elegans	TGGTTGTCATGTCGACG
C54C6.6_Crem_c1	C remanei	CGGTTGTCATGTCGACG
C54C6.6_Cbre_c1	C brenneri	CGGTTGTCATGCCTACG
C54C6.6_Cbri_c1	C briggsae	TGGTTGCCATGTCGACC
C54C6.6_Bmal_c1	B malayi	TAGTTTGCATGACATAA
C54C6.6_Ppac_c1	P pacificus	TGGTTGCCATGGCAAAT

Figure 3.14 – Examples of Applications

A – A 120-bp motif upstream of *Y73B3A.12*, a member of the Calmodulin family; B – A 143-bp motif upstream of *grd-7* (*F46H5.6*); C – A deeply conserved element upstream of *rps-13* (*C16A3.9*) with a table showing motif sequences in all eight species; D – A DAF-19-like site upstream of *kin-2* (*R07E4.6b*); E – An ATF4-like site upstream of *Y34B4A.10*; F – A DAF-19-like site upstream of *C54C6.6* with a table showing motif sequences in four species from genus *Caenorhabditis*, plus *B. malayi* and *P. pacificus*. cisRED URLs are indicated in Table 3.1.

3.7 References

3.7.1 Supplementary Material

Supplementary Material	Link
Supplementary Table 1	http://nar.oxfordjournals.org/content/vol0/issue2008/images/data/gkn1041/DC1/nar-02132-s-2008-File008.xls
Supplementary Table 2	http://nar.oxfordjournals.org/content/vol0/issue2008/images/data/gkn1041/DC1/nar-02132-s-2008-File012.xls
Supplementary Table 3	http://nar.oxfordjournals.org/content/vol0/issue2008/images/data/gkn1041/DC1/nar-02132-s-2008-File014.xls

3.7.2 Works Cited

1. Levine M and Tjian R (2003) Transcription regulation and animal diversity. Nature 424: 147-151.
2. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C *et al.* (2007) WormBase: new content and better access. Nucleic Acids Res 35: D506-10.
3. Okkema PG and Krause M (2005) Transcriptional regulation. WormBook 1-40.
4. Gaudet J, Muttumu S, Horner M, Mango SE (2004) Whole-genome analysis of temporal gene expression during foregut development. PLoS Biol 2: e352.
5. GuhaThakurta D, Schriefer LA, Waterston RH, Stormo GD (2004) Novel transcription regulatory elements in Caenorhabditis elegans muscle genes. Genome Res 14: 2457-2468.
6. Etchberger JF, Lorch A, Sleumer MC, Zapf R, Jones SJ *et al.* (2007) The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. Genes Dev 21: 1653-1674.
7. McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ *et al.* (2007) The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. Dev Biol 302: 627-645.
8. Bulyk ML (2003) Computational prediction of transcription-factor binding site locations. Genome Biol 5: 201.
9. C. elegans Sequencing Consortium (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282: 2012-2018.
10. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR *et al.* (2003) The Genome Sequence of Caenorhabditis briggsae: A Platform for Comparative Genomics. PLoS Biol 1: E45.
11. Sudhaus W and Kiontke K (2007) Comparison of the cryptic nematode species *Caenorhabditis brenneri* sp. n. and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the *Caenorhabditis Elegans* group. Zootaxa 1456: 45-62.
12. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K *et al.* (2008) The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. Nat Genet 40: 1193-1198.
13. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q *et al.* (2007) Draft genome of the filarial nematode parasite Brugia malayi. Science 317: 1756-1760.
14. Mitreva M and Jasmer DP (2006) Biology and genome of Trichinella spiralis. WormBook 1-21.
15. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F *et al.* (2004) Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A 101: 9003-9008.
16. Cutter AD (2008) Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. Mol Biol Evol 25: 778-786.
17. Mitreva M, Blaxter ML, Bird DM, McCarter JP (2005) Comparative genomics of nematodes. Trends Genet 21: 573-581.

18. Kent WJ and Zahler AM (2000) Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. Genome Res 10: 1115-25.
19. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B *et al.* (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 9: 447-64.
20. Robertson G, Bilenky M, Lin K, He A, Yuen W *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res 34: D68-73.
21. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36: D107-13.
22. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: 108-110.
23. Bryne JC, Valen E, Tang ME, Marstrand T, Winther O *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: 102-106.
24. Baillie DL and Rose AM (2000) WABA success: a tool for sequence comparison between large genomes. Genome Res 10: 1071-3.
25. O'Brien KP, Remm M, Sonnhammer EL (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33: D476-80.
26. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B *et al.* (2003) Genome-specific higher-order background models to improve motif detection. Trends Microbiol 11: 61-6.
27. Gaudet J and Mango SE (2002) Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science 295: 821-825.
28. Shostak Y, Van Gilst MR, Antebi A, Yamamoto KR (2004) Identification of C. elegans DAF-12-binding sites, response elements, and target genes. Genes Dev 18: 2529-2544.
29. Hertz GZ and Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563-77.
30. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR *et al.* (2005) Analysis of xbx genes in C. elegans. Development 132: 1923-34.
31. Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 18: 6097-6100.
32. GuhaThakurta D, Palomar L, Stormo GD, Tedesco P, Johnson TE *et al.* (2002) 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: 701-12.
33. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32 Database issue: D91-4.
34. McGinnis S and Madden TL (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res 32: W20-5.
35. Barrière A, Yang S, Pekarek E, Thomas C, Haag ES *et al.* (2009) Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. Genome Res 19: 470-480.
36. Bando T, Ikeda T, Kagawa H (2005) The homeoproteins MAB-18 and CEH-14 insulate the dauer collagen gene col-43 from activation by the adjacent promoter of the Spermatheca gene sth-1 in Caenorhabditis elegans. J Mol Biol 348: 101-112.

37. Baird SE (1999) Natural and experimental associations of Caenorhabditis remanei with Trachelipus rathkii and other terrestrial isopods. Nematology 1: 471-475.
38. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R *et al.* (2003) Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 421: 231-237.
39. Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ *et al.* (2000) Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III. Nature 408: 331-336.
40. Hillier LW, Miller RD, Baird SE, Chinwalla A, Fulton LA *et al.* (2007) Comparison of C. elegans and C. briggsae Genome Sequences Reveals Extensive Conservation of Chromosome Organization and Synteny. PLoS Biol 5: e167.
41. Chen N, Mah A, Blacque OE, Chu J, Phgora K *et al.* (2006) Identification of ciliary and ciliopathy genes in Caenorhabditis elegans through comparative genomics. Genome Biol 7: R126.
42. Haerty W, Artieri C, Khezri N, Singh RS, Gupta BP (2008) Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. BMC Genomics 9: 399.
43. Whetstine JR, Nottke A, Lan F, Huarte M, Smolikov S *et al.* (2006) Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. Cell 125: 467-481.
44. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
45. McKay SJ, Johnsen R, Khattra J, Asano J, Baillie DL *et al.* (2003) Gene expression profiling of cells, tissues, and developmental stages of the nematode C. elegans. Cold Spring Harb Symp Quant Biol 68: 159-169.
46. Schade MA, Reynolds NK, Dollins CM, Miller KG (2005) Mutations that rescue the paralysis of Caenorhabditis elegans ric-8 (synembryn) mutants activate the G alpha(s) pathway and define a third major branch of the synaptic signaling network. Genetics 169: 631-649.
47. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369-73.

4 Conserved Elements Associated with Ribosomal Genes and Trans-splice Acceptor Sites[3]

4.1 Introduction

The cisRED database ([1]; Chapter 3) contains 158 017 conserved motifs in the upstream regions of 3847 *C. elegans* transcripts. These motifs were identified by comparing the *C. elegans* upstream regions to their orthologous counterparts in seven other nematode genomes. Twenty-six percent of the motifs were found to be similar to known transcription factor binding sites (TFBSs) from the ORegAnno [2], JASPAR [3], and TRANSFAC [4] databases. However, the significance and function of almost all motifs remains unknown. We anticipate that many of them may represent previously undiscovered regulatory elements.

Here, we attempted to find novel functional elements by identifying sequences found repeatedly among the cisRED motifs and placing them into groups based on sequence similarity. We assessed the motif groups' significance with respect to the function and expression of their associated genes. Finally we tested the most significant motif group for regulatory function using green fluorescent protein (GFP) and electrophoretic mobility shift assays (EMSAs).

We started with all 158 017 motifs from the cisRED *C. elegans* database of conserved elements (Figure 4.1). These motifs were short sequences from the upstream regions of *C. elegans* protein-coding transcripts that are partially or completely conserved in the orthologous regions of three or more other nematodes. In this analysis we used only the *C. elegans* sequence of each cisRED motif.

We used the DME word-counting motif discovery algorithm [5] to find sequences and variations that were found more frequently in conserved portions of *C. elegans* upstream regions (i.e. cisRED motifs) than in the upstream regions in general. This strategy simultaneously formed groups of motifs that contained the same sequence. We then used the web-based tool DAVID [6] to assess the significance of each motif group by determining whether genes that shared members of the same motif group also shared Gene Ontology, PFAM, and other annotations. We observed that eight of the first 20 motif groups of width-12 base pairs (bp) were significantly associated with ribosomal genes, so we focused further research on these.

The most significant ribosome-associated motif group (the "constitutive motif") was associated with 120 genes, of which 28 were ribosomal, and others were involved in embryonic development, larval development, and multicellular organismal development. For eleven of the

[3] This chapter has been submitted for publication. Monica C. Sleumer, Allan K. Mah, David L. Baillie, and Steven J. M. Jones. **Conserved Elements Associated with Ribosomal Genes and their Trans-splice Acceptor Sites in *Caenorhabditis elegans***

120 genes, we had GFP construct expression data from the BC *C. elegans* Gene Expression Consortium [7]. We tested the importance of the motif for the expression pattern of each of the eleven genes using a series of GFP constructs. Four of the eleven genes showed a difference in expression between constructs including the motif and constructs excluding the motif or with a mutated motif. We then tested the motif from two of these genes for the ability to bind nuclear proteins via an electrophoretic mobility shift assay.

4.2 Results

4.2.1 Motif Grouping

Motifs were formed into 3265 groups based on sequence similarity. DME found sequences that appeared more often than expected in the set of cisRED motif sequences compared to the 3847 upstream regions from which the motifs were derived. We used DME in an iterative process; the most significant motif group was found first, and then the central two bases of each instance of the group were masked with Ns before the next iteration. This way we ensured that each group was unique and not just a minor variation or 1-bp shift of a previously found motif.

DME iterations were run independently for widths of 6, 8, 10, 12, and 14 bp, the same primary widths as the cisRED motifs themselves (Table 4.1). All 158 017 of the motifs in the cisRED database had a width of at least six bp, therefore they were all eligible for grouping at that width; in contrast, only 72 935 of the cisRED motifs had a width of at least 14 bp.

The parameters for DME were set such that the stringency for motif sequence similarity was relaxed as the motif width increased. The requirements for width-six motif groups were set to a maximum stringency: an information content (IC) of two or perfect match. This meant that all members of motif groups of width six contained the same hexamer. For motif groups of width eight, the IC requirement was 1.8, meaning that each motif group consisted of motifs containing octamers that differed in only one position. For motif groups of width 14, the IC requirement was only 1.5. In spite of the relaxed IC requirements, motif groups of high width were far smaller than those of low width. This was due to the relative rarity of long sequences, and the smaller set of motifs from which to draw.

Seventy-six overrepresented hexamers were found among the cisRED motifs. The first motif group was the hexamer GATAAG. This sequence appeared 1469 times among the *C. elegans* sequences of the cisRED motifs and only 2245 times among all *C. elegans* upstream regions from which the cisRED motifs were derived. Because the cisRED motif sequences amounted to a total of 2.15 Mbp and the upstream regions amounted to a total of 4.16 Mbp (nonrepeat-masked), the expected number of occurrences for the hexamer among the motifs was 1160, therefore this hexamer was significantly over-represented. Similarly, the second hexamer motif group consisted of motifs containing the palindrome CACGTG. This sequence appeared 699 times among the *C. elegans* upstream regions, of which 502 were conserved. After 76 hexamer-based motif groups were found, DME was no longer able to find any further hexamers that were over-represented among the cisRED motifs with respect to the upstream regions. The 76th hexamer motif group was the sequence GCGTTG, which appeared 813 times in total among the upstream regions and 412 times among the cisRED motifs, only slightly more often than the expected number of 407. The smallest hexamer motif group was the palindrome GGGCCC, which appeared 65 times among cisRED motifs and 118 times in the upstream regions. After 76 iterations, 28 478 (18%) of motifs were in at least one hexamer motif group (Table 4.1).

Four hundred eighty-nine overrepresented octamers were found. Of the cisRED motifs, 11 965 were of width 6 or 7 (Figure 3.6) and hence were ineligible for width-eight motif grouping, leaving 146 052 available motifs. The first motif group of width 8 consisted of motifs that contained the sequence CTGCGYCT. This sequence appeared 506 times in total among the upstream regions and 371 times among cisRED motifs (the expected number of occurrences was 261). Members of the second octamer motif group all contained the sequence GHGCGCGC. This sequence is a partial palindrome (whenever the base in the second position is a C). It is also possible for instances of this sequence to overlap substantially with each other. Including all overlapping instances on both strands, DME counted a total of 447 instances of this sequence among the upstream regions, of which 328 were in cisRED motifs. The smallest octamer motif group was the palindrome GGCTAGCC, which appeared twice among cisRED motifs and only three times in total among the upstream regions.

After 489 motif groups were found and their central bases were masked out, DME was unable to find any more octamers that were overrepresented among cisRED motifs. The last octamer motif group found was TGACGGTG, which appeared 73 times in the upstream regions, of which 37 were in cisRED motifs; the expected number of occurrences among cisRED motifs was 36. Many of the octamer motif groups overlapped with each other and with the hexamer motif groups; 19 824 of the motifs were in at least one octamer motif group.

Nine hundred overrepresented decameric motif groups were found. Of the entire set of cisRED motifs, 125 822 were at least 10 bp wide and therefore eligible for decameric grouping. The first width-10 motif group had the consensus sequence GAKACGCAGN; sequences that matched this pattern appeared 402 times in the upstream regions, of which 282 were within cisRED motifs. The second decameric motif group had the consensus sequence GTCYCGCMRC, which appeared in 155 cisRED motifs and 215 times in the upstream sequences. The smallest decamer motif group was the palindrome CAATGCATTG, which appeared twice among the upstream regions; both occurrences were in cisRED motifs. DME did not terminate automatically and was stopped after 900 iterations had run. The 900th decameric motif group had the consensus sequence TAACMCGWCT, which appeared 14 times in the upstream regions, 11 of which were in cisRED motifs. After 900 iterations, 16 583 of the motifs were in decameric motif groups.

Nine hundred overrepresented dodecameric motif groups were found. There were 101 348 cisRED motifs with a width of 12 or more bp. Many of the first few width-12 motif groups were very interesting and are discussed in detail below. The two smallest dodecameric motif groups were the palindromes MTACTRYAGTAK and RWGTTGCAACWY. Both of these sequences had five occurrences in the upstream regions, all of which were in cisRED motifs. Just as for the decameric motif groups, DME did not terminate automatically and was stopped after 900 iterations; The last dodecameric motif group had the consensus sequence YGGCGGCRSCAB. Sequences matching this pattern were observed 12 times in the upstream regions and 11 times among cisRED motifs. After 900 iterations, 11 963 of the motifs were in dodecameric motif groups.

Nine hundred overrepresented width-14 motif groups were found. Only 72 935 (46%) of the motifs were wide enough to be eligible for width-14 motif grouping. The first width-14 motif group overlapped very strongly with the first width-12 group, with a consensus sequence of KSGTCYSSSMRCGA. However, it was not a superset, because the width-12 group included some motifs that were exactly 12 bp wide, and it was also not a subset because the IC of the width 14 group was lower and more sequence variations were included. The second group had the consensus sequence RYRWGTGYKASYGT, which appeared 44 times in the upstream regions and 37 times among the cisRED motifs. The smallest motif group had the consensus sequence SWGCCWYRWGGCWS, which appeared five times in the upstream regions, four of which were in cisRED motifs. After 900 iterations, DME was terminated. The last motif group found had the consensus sequence RWMAWTMTYGKCGT which appeared six times among the upstream regions, all of which were in cisRED motifs. Of the eligible motifs, 8725 of them were in groups after 900 iterations.

In total, 45 312 (29%) of motifs were in 3265 overlapping groups after all DME iterations were completed. Half of the motifs were in more than one group. A summary of the motif grouping result numbers is shown in Table 4.1. Motif groups are browsable via the cisRED web interface at www.cisred.org/c.elegans4/all_groups?showab=0&showdn=1. Additionally, each motif group has its own URL at www.cisred.org/c.elegans4/group_content_view?aid={Group ID} (note that "{Group_ID}" must be replaced by the Group ID of the motif group in question).

4.2.2 Functional Characterization of Genes with DAVID

We used DAVID [6] to examine the associated genes of each of the first 20 motif groups at each width to see if they had any functional similarities. For the hexameric motif groups, all 20 groups had significant multiple testing-corrected associations. For example, the fifth group, consisting of all instances of the sequence GGGCGG among the motifs, was significantly associated with genes involved in DNA binding, including homeobox genes. Six of the motif group associations were with ribosomal proteins.

For the octameric motif groups, 17 out of the first 20 groups had significant associations. For example, the first motif group was associated with ATP-binding and mitochondrial proteins. Four of the top 20 motif groups were associated with ribosomal proteins. Similarly, 16 of the first 20 decameric motif groups were significantly associated with gene categories such as nucleotide binding, cytoplasmic proteins, transit peptides, and anatomical structure development. Once again, eight of the decameric groups were associated with ribosomal proteins. The same general results were observed for dodecameric and width-14 motif groups: 11 of the dodecameric groups were significant (eight of them with ribosomal proteins), and eight of the width-14 groups were significant, six of them with ribosomal proteins.

4.2.3 Description of Eight Motif Groups Associated with Ribosomal Genes

We observed that many of the significant motif groups were associated with ribosomal genes, so we decided to concentrate further research on these motifs. Specifically, eight of the top 20 dodecameric motif groups were associated with between six and 28 ribosomal genes, with a total of 63 ribosomal genes between them (Table 4.2). There were only 96 ribosomal genes in the set of genes used for the cisRED database, and there are only 176 ribosomal protein genes in total in the *C. elegans* genome, so this was a substantial proportion. Three pairs of ribosomal proteins were on bidirectional promoters and therefore had the same upstream region. Each ribosomal gene had no more than one instance of each motif group (with one exception: there were two instances of motif group 1469 upstream of *Y119D3B.16*) and no more than three different motif groups in its upstream region.

In order to determine whether any of the motifs were similar to previously determined TFBS in nematodes and other organisms, we first looked at whether they were associated with cisRED annotations (section 3.2.5). The results for the first motif group (hereafter referred to as the "Constitutive Motif") are described below. Fifteen of the 35 motifs in group 1474 were annotated as similar to the UNC-30 site, which is TAATCC [8]; none of the other motifs had any significant association with cisRED motif annotations.

Some of the motif groups were similar to known TFBS in other species. We then performed a second comparison using the Transcriptional Element Search System (TESS) [9]. Once again results for the Constitutive Motif are described below. Motif Group 1467 was found to be significantly similar to the binding site for mouse ZF5, whose consensus binding sequence is GSGCGCGR [10] (Table 4.2). Motif group 1469 had significant similarities to the binding sites for both EBP-45 (binding sequence TGTTTGC [11]) and HNF3-family transcription factors (binding consensus sequence described as YGTTTRT in rat [12] and TRTTTGY in the frog *Xenopus laevi*s [13]). Motif group 1474 was found to be significantly similar to the binding site for Delta EF1 in the chicken genome (binding sequence AGGTG [14]) even though the motif group sequence was not a perfect match. Motif group 1484 was significantly similar to the Zeste binding site in *Drosophila melanogaster* (binding consensus sequence YGAGYG [15]).

The motif groups are associated with cytoplasmic ribosomal proteins. *C. elegans*, like other non-photosynthetic eukaryotes, has ribosomes associated with two different intracellular localization patterns, cytoplasmic and mitochondrial [16]. We examined whether the ribosomal motif groups were associated with cytoplasmic ribosomal genes, mitochondrial ribosomal genes, or both. Mitochondrial ribosomal genes are not specifically annotated in the *C. elegans* genome, but some of them are tentatively identified with KOG (eukaryotic clusters of orthologous groups) designations [17]. Of all ribosomal transcripts in the cisRED database, 102 were annotated as ribosomal by KOG, and 24 of these were annotated as mitochondrial ribosomal proteins. Of the 63 ribosomal genes with a ribosomal motif in their upstream region, 50 were annotated as ribosomal by KOG, and only three of these were annotated as mitochondrial ribosomal (*B0303.15*, *K07A12.7*, and *Y48C3A.10*). The two-tailed p-value for this distribution is 4.2e-5 by the Fisher Exact Test; therefore, the motifs we found are specifically associated with cytoplasmic ribosomal genes and not with mitochondrial ribosomal genes.

4.2.4 Motifs Overlapping Trans-splice Sites

Motif group 1474 is an extension of a trans-splice acceptor site. We observed that one of the motifs (Group ID 1474) had a strong strand bias: 28/35 instances were on the same strand as the nearest gene. We also observed that this motif almost always occurs about 20bp upstream of the translation start site (ATG). Those instances on the opposite strand tended to be on bidirectional promoters, at the far end of the upstream region, and therefore just upstream of (and on the same strand as) the other gene of the promoter. We also found that 28/35 instances of this motif group overlapped annotated SL1 and SL2 trans-splice sites in Wormbase.

Several other motif groups are extensions of trans-splice acceptor sites. In order to investigate the connection between motif groups and trans-splice acceptor sites more thoroughly, we searched for other motif groups that were also associated with trans-splice acceptor sites (Table 4.3). Of all 3265 motif groups, at least 16 had a majority of motifs that overlapped with trans-splice sites. Almost all of these also had significant associations with ribosomal genes (some were too small in number to have significant associations). All were variations or extensions of the canonical trans-splice site TTTCAG [18].

4.2.5 Motif Group 1466 – The Constitutive Motif

The first dodecameric motif found by DME was the largest and the most significant with respect to ribosomal genes, so we investigated further to try to determine its function. We observed that it was GC-rich and very strongly conserved, especially the instances near ribosomal genes. It tended to appear about 300bp upstream of the ATG of the gene and was not strand-biased (Figure 4.2). It also tended to be found about 30bp upstream or downstream of one of the other ribosomal motifs such as 1471, 1477, or 1484. It was not found to co-occur in an upstream region with motif group 1469 or 1470.

The constitutive motif was found upstream of 28 ribosomal genes, of which two pairs of genes were on bidirectional promoters and therefore had the same upstream regions: *lsm-1* (*F40F8.9*; a small nuclear ribonucleoprotein splicing factor) and *rps-9* (*F40F8.10*), and *rps-30* (*C26F1.4*) and *rpl-39* (*C26F1.9*).

The constitutive motif was not similar to a known TFBS. We noticed that 89/147 of the motif instances were annotated in cisRED as similar to the HSAS element described by GuhaThakurta *et al.* [19]. Specifically, GuhaThakurta *et al.* showed that the sequence GGGTCTC was involved in the regulation of *hsp-16-2*; the subsequence GGTCTC (reverse complement of GAGACC) was part of the constitutive motif. The result was that all instances of this motif that had a further C at the end of the motif were annotated as similar to the HSAS element, accounting for 25% of all cisRED motifs annotated as similar to the HSAS element. However, none of the HSAS elements in the GuhaThakurta *et al.* paper bore any further resemblance to the constitutive motif. The rest of the constitutive motif is GC-rich and the HSAS elements tended to be flanked by AT-rich sequence, so it is unlikely that they are connected. We also used TESS to search for any other similarities to known TFBS, but the search returned no results of interest.

4.2.6 GFP Testing of Function

We performed a series of GFP experiments to determine whether the presence of motif was related to gene expression. The BC *C. elegans* Gene Expression Consortium had previously created many GFP constructs and recorded their subsequent expression pattern [7]. Those constructs were made using 3 kbp upstream regions or the intergenic region if it was less than 3 kbp. The focus of the Consortium was on genes with human orthologues, but very few ribosomal genes were included. Of the 120 genes with an instance of motif group 1466 in their upstream regions, 11 had had GFP constructs made by the Consortium, only one of which was a ribosomal gene. However, all 11 of these constructs drove strong expression across a number of different tissues and stages of development, and were therefore good candidates for further dissection of their promoter regions for assessment of promoter activity.

For each of the eleven upstream regions that had both an instance of the constitutive motif and previous GFP expression data, three GFP constructs were made: one that included the motif, one that excluded the motif, and one that introduced a mutation in the center of the motif (Figure 4.3). These constructs were injected into the gonad of gravid hermaphrodites, and the worm progeny were allowed to grow to adulthood. Photographs were taken of the worms and their GFP expression was observed and recorded.

GFP expression constructs indicated that the constitutive motif is involved in regulation of pharyngeal expression. For four of the genes, we found that the construct including the motif produced some GFP expression in the pharynx while the construct excluding the motif and the construct with a mutated motif showed little or no pharyngeal expression (Table 4.4, “Tentative Positives”, Table 4.5). Two of the genes had inconclusive results because the GFP expression was not correlated with the presence of the motif in the construct (Table 4.6). Two genes showed no difference in gene expression between the three constructs and were therefore construed as negative results with respect to motif function. Three genes had no GFP expression at all from any of the three constructs, and therefore the function of the motif can not be determined.

4.2.7 Electrophoretic Mobility Shift Assay

We tested two of the primers from the GFP experiments done on *C34E10.6* and *F25H2.5* to see if protein binding could be detected via an electrophoretic mobility shift assay. We used a biotinylated probe and compared the bands resulting from the free probe, probe with cytosolic extract, and probe with nuclear extract. We also performed a competition assay using varying concentrations of unlabelled competitor probes, both the same-sequence and mutated.

The electrophoretic mobility shift assay did not indicate protein binding to the constitutive motif. None of the gel lanes showed any shifted bands (Figure 4.4). There was no difference in the results from the free probe lanes, the lanes containing nuclear extract, and the lanes containing unlabelled competitor. The positive control did show a shifted band in the lane containing nuclear extract.

4.3 Discussion

The DME iterative process, while computationally inefficient, was very effective in identifying sequences that were conserved in the upstream regions of *C. elegans* protein-coding transcripts more often than expected (Table 4.1). We observed that it tended to skew towards relatively GC-rich sequences because the AT content was considerably lower among cisRED motifs (60.7%) than among the upstream regions (65.8%). It also found the largest group first; there was a general trend towards smaller and smaller groups as the DME iterations progressed.

DME counted each instance of a motif group including overlapping sequence instances; palindromes were counted twice. This meant that DME skewed slightly towards repeating and palindromic sequences: the total counts were higher and therefore the difference between the foreground count and the expected foreground count was higher. We did not consider this to be a confounding factor because TFBSs are sometimes palindromic or partially palindromic due to the binding of homodimeric TFs. For example, the *C. elegans* X-box TF DAF-19 binds an imperfect palindromic sequence [20].

A confounding issue was that many of the cisRED motifs overlapped substantially. In a few cases a series of overlapping cisRED motifs caused DME to identify a sequence as over-represented when most or all instances in the foreground referred to a single genomic location. However, some of the upstream regions overlapped as well – bidirectional promoters and alternative transcripts of the same gene – which mitigated the effect of overlapping cisRED motifs somewhat.

We observed that DME found motifs that appear more often in the foreground than the background, but not necessarily significantly more often. For example, the last octamer motif group appeared 73 times in the upstream regions, of which 37 were in cisRED motifs. The expected number of occurrences among cisRED motifs was 36. By the Fisher Exact Test, the p-value for this distribution is 0.57 because the experimental value is so close to the expected value. In contrast, the first octamer motif group appeared 506 times in the upstream regions and 371 times among cisRED motifs, and the expected number was only 261; the p-value for this distribution is less than 10^{-100}. This is why we terminated DME after 900 iterations for motif widths 10, 12, and 14; the low information content requirement made it possible for DME to find a virtually unlimited number of very small motif groups of dubious importance and significance.

An advantage that DAVID has over other Gene Ontology (GO) [21] analysis tools is that it is able to determine whether gene groups are enriched for terms from other gene annotation sources such as the Protein Information Resource [22] and the KEGG Pathway Database [23] in addition to the GO itself. We found that the PIR keywords tended to be both more specific than the GO terms and had annotations for more of the genes associated with motif groups, and as a result we obtained more information about the motif groups than we would have from looking at only GO terms.

We also found that DAVID has several disadvantages. It is a web-based tool that is not designed to be used in a high-throughput way. The HTML-based API is limited both by the maximum URL length and by the internal limit of 400 genes – several of our gene groups exceeded this limit and were not analyzed completely by DAVID. Although it is possible to upload a background gene list for a single gene list, it is not possible to use the correct background list in the API, with the result that some of our significant p-values may be off by several orders of magnitude. However, due to the extreme p-values, it is not expected that this incongruity impacted the true significance of any of the motif groups described here.

Any gene group enrichment analysis method will produce some false negatives. A lack of significant associations does not mean that the genes have nothing further in common. The possibility of gene group significance decreases as the group size decreases; small gene groups (less than ten genes) will not be significant unless all of them fall into a specific and rare category. Large gene groups will be highly significant even when only a minority of the genes fall into the same category.

Eight of the top 20 motif groups in both decameric and dodecameric series of DME iterations were enriched for ribosomal genes. We decided to concentrate our efforts on the dodecameric motif groups because although both sets had similar p-values and were probably equally valid, the dodecameric groups had fewer genes associated with them, making further analysis more straightforward.

Several of the ribosomal motif groups were similar to experimentally validated TFBSs as determined by TESS (Table 4.2). Motif group 1467 was found to be similar to mouse ZF5, but the section of the motif group that matched ZF5 was also the least-conserved portion, so it is unlikely that the similarity is important. The similarity of motif group 1467 to EBP-45 and HNF3-family TFs may be more interesting. Both of these TFs have similar binding sites, and because HNF3-family TFs are conserved between frog and rat, they may also be present in *C. elegans*. The similarity of motif group 1474 to the rat Delta EF1 site is probably not important. The site sequence is not a perfect match, and it seems likely that the importance of this site and the other trans-splice acceptor site-related motif groups is due to splicing factors binding to the RNA, not TFs binding to the DNA. The similarity of motif group 1484 to the *Drosophila* Zeste site is worthy of note: the site similarity is in the perfectly-conserved portion of the motif group. Additionally, Zeste is a polycomb-group protein that has a known orthologue in *C. elegans*: MES-2 [24]. The *C. elegans mes-2* knockout mutant has the maternal effect sterile phenotype, which means that MES-2 is an important protein required for germline development.

Closer inspection of the strand- and location-biased motif group 1474 revealed that not only was it an extension of a trans-splice acceptor site, but also that several other motif groups were also trans-splice acceptor site extensions (Table 4.3). It makes sense that the trans-splice locations would be conserved in the orthologous regions, as it is logical that the other nematodes also perform trans-splicing of transcripts. The canonical trans-splice site is TTTCAG; our results suggest that the trans-splice acceptor site may be more complex. One of the motif groups (1082) was a noncanonical extension at the 5' end of the trans-splice site: for nine genes (of which three were confirmed ribosomal), we saw the pattern GTAATCCAG at the trans-splice site. The other motif groups were all extensions of the trans-splice site at the 3' end, beyond the CAG. There were three specific extensions of the pattern: CAGGTAA (motif groups 87, 569, 1474, and 2376), CAGGGTA (motif groups 111 and part of 580), and CAGGGTT (motif groups 365 and part of 580). It is not clear why ribosomal genes in particular would have special trans-splice acceptor sites. Perhaps ribosomal transcripts have a signal that fast-tracks them for processing and translation, allowing other transcripts to be translated after more ribosomes are made.

Although it is clear that the ribosomal genes discussed here are in general enriched for cytoplasmic ribosomal genes, it is possible that one or more of the motif groups is associated with nuclear-encoded mitochondrial ribosomal genes. The three KOG-annotated mitochondrial ribosomal genes only had instances of motif groups 1469 and 1470 in their upstream regions. These same two motif groups were found to co-occur the least with other ribosomal motif groups. In contrast, half of the instances of the constitutive motif (group 1466) were found to occur in close proximity to instances of motif groups 1471, 1477, or 1484 in a striking pattern (Figure 4.2). Without more information as to the specific function of each of these ribosomal genes, it is difficult to investigate these occurrence patterns in more depth.

Four of the eleven genes tested using GFP constructs displayed a dependence on the constitutive motif for pharyngeal expression (Table 4.4, Table 4.5). It is not clear why the motif was related to expression in the pharynx as opposed to other tissues, because these genes are normally expressed in most or all tissues. For one of the positive results (*F09B9.3*), the constitutive motif was in the WormBase-annotated 5' UTR, suggesting that the genome contains an additional transcription start site between the motif and the gene's ATG. The motif seemed to be better-conserved in the upstream regions of the genes that had positive indications of function, but most of them were well enough conserved to be found repeatedly by motif discovery of the upstream regions and their orthologues. The motif was very poorly conserved for two of the eleven genes (*Y57G11C.13* and *T05H4.1*, which had "unclear" and "tentative negative" results respectively). They were so poorly preserved that although they were found within cisRED motifs in an earlier unpublished version of the cisRED database, they were not within motifs in the published cisRED database. Only one ribosomal gene was tested with GFP (*rpl-17*, or *Y48G8AL.8a*), but because no GFP expression was observed for any of the three constructs, we were unable to determine whether the motif was involved in the regulation of the expression of this gene.

The electrophoretic mobility shift assay did not display any evidence of protein binding to the constitutive motif *in vitro*. There are two possible explanations for this result: either the motif does bind a TF and we were unable to detect it, or the motif does not bind a TF. If we were unable to detect TF binding that does occur *in vivo*, it is possible that the concentration of the TF in the nuclear extract was too low for us to be able to see a shifted band. It is also possible that the DNA sequence we used for the assay was too short for the protein to bind, or that the constitutive motif only binds a TF in concert with another nearby sequence such as one of the other motifs that we found. If the motif does not bind a protein *in vivo*, it must have some other function because its conservation is too significant to have occurred by chance, and its association with ribosomal genes is very strong. It may be part of the 5' UTR in all of the genes and thereby functions at the RNA level rather than the DNA level, for example as an RNA binding protein binding site or antisense RNA binding site.

4.3.1 Conclusions

The motif grouping program DME was successful in finding interesting sequences that were conserved in the orthologues much more often than expected. The motif groups had significant functional associations, showing that the repeated, evolutionarily conserved sequences that we found could not have occurred by chance and have some sort of biological importance. The p-values for the ribosomal motif groups were extremely low after multiple testing correction was performed, and robust in the sense that similar statistics were calculated repeatedly, regardless of variations in the width of the motif and the IC.

At least one of the eight ribosomal motifs is similar to a known binding site of a TF that has a *C. elegans* orthologue and warrants further investigation of this connection. Trans-splice sites are strongly conserved for ribosomal genes and follow specific patterns that are extensions of the canonical trans-splice sites.

The constitutive motif is usually found 300 bp upstream of the ATG of ribosomal genes and tends to occur in close proximity to instances of motif groups 1471, 1477, or 1484. GFP construct experiments in broadly expressed genes indicated that it may have a direct impact on the pharyngeal expression of some genes, but its influence on the expression of ribosomal genes remains undetermined. Electrophoretic mobility shift assays were inconclusive and did not determine whether the the motif requires flanking sequence to bind a protein detectably or whether the motif functions at the RNA level rather than the DNA level.

4.4 Methods

The *C. elegans* sequence of all cisRED motifs were extracted and combined into a single FastA file. All *C. elegans* cisRED upstream regions were combined into a background FastA file (original genome build: WS170). A version of DME that did not preface the word-counting step with a repeatmasking step and did not weight motif IC by base composition was obtained from Dr. Andrew Smith. DME was run using the parameters indicated in Table 4.7 at each width. After each iteration, the two central bases of each motif in the new group were masked to Ns and DME was re-run until it either could no longer find any overrepresented sequences, or reached 900 iterations. All motif groups were uploaded to the cisRED database as ‘de novo’ motif groups.

Entrez gene IDs for all genes associated with the first 20 groups at each width were extracted and analyzed by DAVID via the HTML-based API. The following annotation categories were included in the HTML links: GOTERM_BP_ALL, GOTERM_CC_ALL, GOTERM_MF_ALL, INTERPRO,PFAM, PIR_SUPERFAMILY, KEGG_PATHWAY, SP_PIR_KEYWORDS, BIND DIP, and MINT. For motif groups with more than associated 400 genes, only the first 400 genes were analyzed (in alphabetical order).

Primers were designed for GFP constructs as shown in Table 4.8. Constructs were generated by PCR and injected into the gonads of gravid hermaphrodites. Worms were incubated at 21 degrees Celsius and progeny were observed and photographed in the microscope five days later.

The "Primer Including Motif" DNA sequence was also used for the electrophoretic mobility shift assays, and the "Primer Mutating Motif" DNA sequence was used as the mutated competitor. Labelled probes were biotinylated at the 5' end of both strands. Primers were annealed by incubating the mixed oligos at 1000x concentration at 65 degrees Celsius followd by slow cooling. Cytosolic and nuclear extracts were obtained from mixed-stage worms. The reactions were run on an ice water-immersed 8% gel in 0.5X TBE. Poly dI-dC was used as a non-specific competitor (as opposed to salmon sperm DNA).

4.5 Chapter 4 Tables

Width (bp)	# of Available Motifs	Min IC	# Groups	Smallest	Largest	# Motifs in Groups	Group ID Range
6	158 017	2.0	76	65	1452	28 478	1 to 76
8	146 052	1.8	489	2	543	19 824	77 to 565
10	125 822	1.7	900	4	282	16 583	566 to 1465
12	101 348	1.6	900	5	155	11 963	1466 to 2365
14	72 935	1.5	900	5	91	8725	2366 to 3265

Table 4.1 – Summary of Motif Grouping Results

DME was run iteratively on cisRED motifs to form them into groups based on sequence similarity. For width 6 bp, all motifs were eligible, while for width 14bp, only 72 935 motifs were available. For widths 6 and 8, DME terminated automatically with no motif groups left to find after 76 and 489 iterations respectively, while for widths 10, 12, 14, the process was terminated after 900 iterations. The number of motifs in each group varied widely between groups.

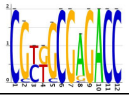
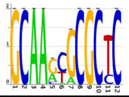
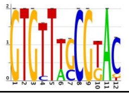
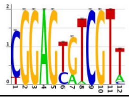
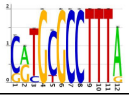
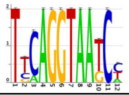
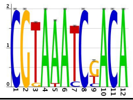
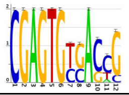
Group ID	Iteration Number	Background Count	Num Motifs	Num Genes	Num Ribosomal	Benjamini P-value	Logo	Characteristics
1466	0	200	147	120	28	2.60E-24		Constitutive Motif; Discussed further below
1467	1	162	113	103	8	2.80E-08		Similar to ZF5 site
1469	3	118	87	76	10	2.40E-11		Similar to HNF3 family TFBS and EBP-45 site
1470	4	86	69	65	6	6.40E-07		
1471	5	99	74	65	14	2.80E-16		
1474	8	36	35	23	8	1.20E-10		Strand bias; Trans-splice site; Similar to Delta EF1 site
1477	11	123	78	63	12	5.70E-14		
1484	18	31	28	21	7	9.30E-04		Similar to *Drosophila* Zeste site

Table 4.2 – Summary of Ribosomal Protein-associated Motif Groups

The first column shows the Group ID of each motif group in the cisRED database, and the second column shows the iteration number of the dodecameric series of motif groups. “Background Count” shows the number of instances of the motif group sequences among all cisRED upstream regions, and “Num Motifs” shows the number of instances of the motif group among cisRED motifs. “Num Genes” shows the number of different genes associated with each motif group, and “Num Ribosomal” shows how many of these were annotated as ribosomal by DAVID, while “Benjamini P-value” indicates the Benjamini-corrected p-value of this proportion of ribosomal genes. Lastly, the logo of each motif group (from all instances, not only ribosomal instances) and other characteristics of each motif group are shown.

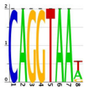
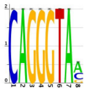
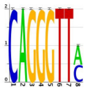
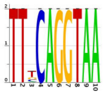
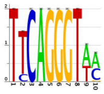
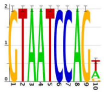
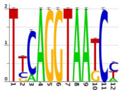
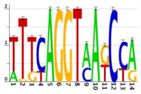
Group ID	Width	Iteration Number	Background Count	Num Motifs	Num Trans-Splice Sites	Num Genes	Num Ribosomal	Benjamini P-Value	Logo
87	8	10	339	224	123	154	19	1.7E-12	
111	8	34	110	76	55	49	11	5.7E-14	
365	8	288	87	47	24	37	9	5.6E-5	
569	10	3	161	119	101	78	21	5.3E-21	
580	10	14	89	66	58	41	13	4.3E-18	
1082	10	516	18	14	11	9	3	6.4E-2	
1474	12	8	36	35	28	23	8	1.2E-10	
2376	14	10	23	22	14	15	4	4.8E-3	

Table 4.3 – Selection of Motif Groups that Overlap Trans-splice Acceptor Sites

The first column shows the Group ID of the motif group in the cisRED database. The second and third columns show the width of the motif group and the DME iteration number for that width. “Background Count” shows the number of instances of the motif group sequences among all cisRED upstream regions, and “Num Motifs” shows the number of instances of the motif group among cisRED motifs. “Num Trans-Splice Sites shows how many of the motifs overlap trans-splice acceptor sites in WormBase. “Num Genes” shows the number of different genes associated with each motif group, and “Num Ribosomal” shows how many of these were annotated as ribosomal by DAVID, while “Benjamini P-value” indicated the Benjamini-corrected p-value of this proportion of ribosomal genes. Lastly, the logo of each motif group (from all instances, not only ribosomal instances) is shown.

ID	Gene Name	Orig. GFP Construct		Construct Incl. Motif		Construct Excl. Motif		Construct w/ Mut. Motif	
Expression:		Pharynx	Other	Pharynx	Other	Pharynx	Other	Pharynx	Other
Motif		+		+		-		Mutated	
Tentative Positives									
A3	*C34E10.6*	+++	+++	**+**	+	**-**	-	**-**	-
A5	*F09B9.3*	+++	+++	**+**	-	**-**	-	**-**	-
A6	*F25H2.5*	+++	+++	**+++**	+	**+**	+	**-**	+
A7	*F54D8.2*	+++	+++	**+++**	++	**+**	+	**-**	+
Unclear									
A2	*C26D10.2*	+++	+++	++	++	+	+	-	-
A11	*Y57G11C.13*	+++	+++	++	++	-	-	+	+
Tentative Negatives									
A4	*F07A11.2a*	+++	+++	++	++	++	++	++	++
A9	*T05H4.1*	+++	+++	++	++	++	++	++	++
No Expression									
A1	*C13B9.3*	+++	+++	-	-	-	-	-	-
A8	*M01F1.3*	+++	+++	-	-	-	-	-	-
A10	*Y48G8AL.8a*	+++	+++	-	-	-	-	-	-

Table 4.4 – Summary of Observed GFP Expression

GFP expression is described for four GFP constructs for each of the eleven genes tested in this study: the expression observed by the BC *C. elegans* Gene Expression Consortium (“Orig. GFP Construct”), from the first construct (“Construct Incl. Motif”), from the second construct (“Construct Excl. Motif”), and from the third construct (“Construct w/Mut. Motif”). GFP expression is separated into pharyngeal expression and expression in all other tissues because pharyngeal expression showed the greatest differences. The level of GFP is indicated by one to three ‘+’, while no GFP expression is indicated by ‘-’. Genes are sorted into four categories: those that showed a clear difference in expression that correlated with the presence of the motif (“Tentative Positives”), those that showed a difference in expression that was not correlated with the presence of the motif (“Unclear”), those that showed no difference in expression between the three constructs (“Tentative Negatives”), and those that showed no GFP expression from any of the three constructs (“No Expression”). Bold values show the loss of pharyngeal expression for the four tentative positive results.

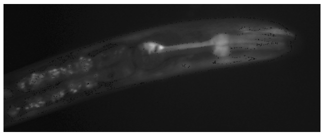
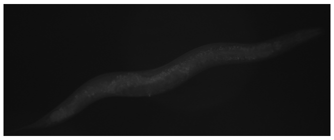
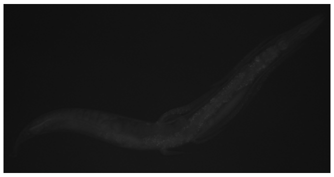
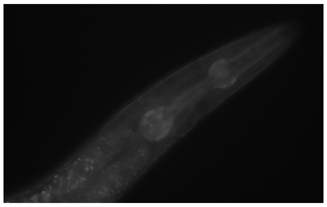
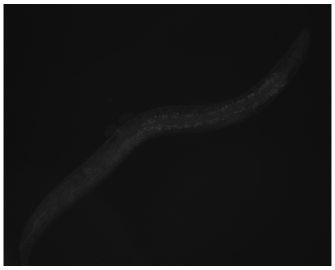
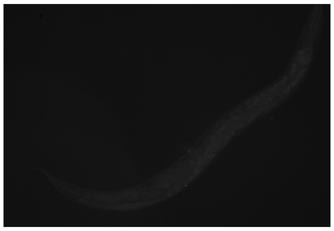
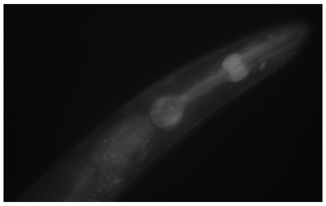
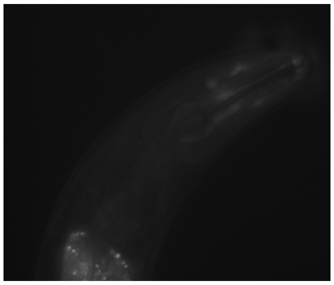
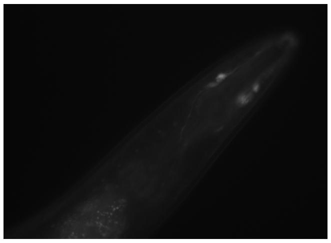
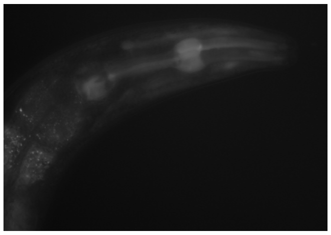
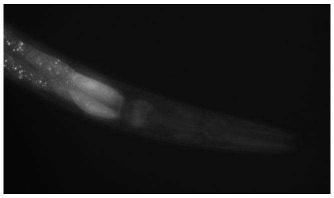
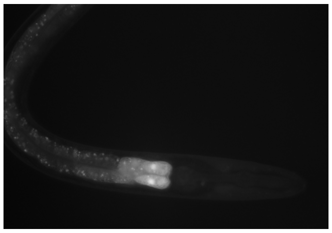
Gene	Construct Including Motif	Construct Excluding Motif	Construct with Mutated Motif
C34E10.6			
F09B9.3			
F25H2.5			
F54D8.2			

Table 4.5 – GFP Images for Tentative Positives

GFP images for the four upstream regions that resulted in a tentative positive indication of motif function. For each upstream, the construct including the motif produced GFP expression in the pharynx, while the constructs excluding the motif and with a mutated motif produced no pharyngeal expression.

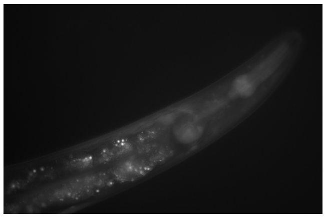
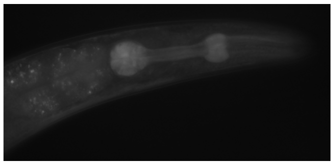
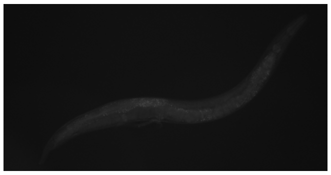
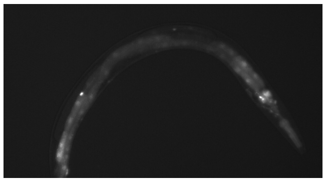
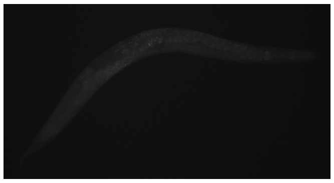
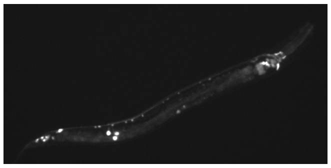
Gene	Construct Including Motif	Construct Excluding Motif	Construct with Mutated Motif
C26D10.2			
Y57G11C.13			

Table 4.6 – GFP Images for Unclear Results

GFP images for the two upstream regions that resulted in an unclear indication of motif function. For each upstream, the construct including the motif produced GFP expression in the pharynx. For *C26D10.2*, the construct excluding the motif also produced some pharyngeal expression and the construct with a mutated motif produced no expression. For *Y57G11C.13*, the construct excluding the motif produced no expression and the construct with a mutated motif produced GFP expression in a variety of tissues.

Width	Parameters
6	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.9 -w 6 -r 0.25 -g 0.0
8	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.8 -w 8 -r 0.25 -g 0.0
10	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.7 -w 10 -r 0.25 -g 0.0
12	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.6 -w 12 -r 0.25 -g 0.5
14	-C 0.25,0.25,0.25,0.25 -n 1 -i 1.5 -w 14 -r 0.25 -g 1.0

Table 4.7 – DME Parameters

Parameters used for the word-counting motif discovery algorithm DME at each of the five widths are indicated.

Gene	Primer Including Motif	Primer Mutating Motif	Primer Excluding Motif	Reverse Primer
C13B9.3	cggggaggtctcgcaacgaaatga	cggggaggtctcttaacgaaatga	ttcactggttgttcgttgga	cggcgatcaacacgattg
C26D10.2	ttacttcgctgcgagaccatacgaa	ttacttcgctaagagaccatacgaa	cgaatgggtatcgtttcgc	gttgttcctcttcgattctgaaa
C34E10.6	tccatttcgttgcgagacccgctg	tccatttcgttaagagacccgctg	gcggtctagcctgtttcagt	taacgaacgcgaagcgata
F07A11.2a	tctcaaccggagcgttgcgagacc	tctcaaccggagcgttaagagacc	tgatctttcgatcgttctcg	aattccgcagattttggatg
F09B9.3	agacgaacatcgctgcgagaccag	agacgaacatcgctaagagaccag	ggacgaatagctcgcatctc	tctgcgttatggaagaacagaa
F25H2.5	aggtcgggtctcgccacgtgctgaagta	aggtcgggtctcttcacgtgctgaagta	tcgtttcatttgtgtcggag	tcagtgttgctgattttcgg
F54D8.2	atttcaccggctggtctcgcagcgaa	atttcaccggctggtctcttagcgaa	agacggcctctccgttattt	cggttgatgtcggatacctt
M01F1.3	attgcgtatcgtggcgagacccat	attgcgtatcgtgaagagacccat	atggctttttccgctatcct	acccgagctaggatgcttaaa
T05H4.1	acttcctgagcgttgcgagacctgt	acttcctgagcgttaagagacctgt	tccacaaaagaacacctccc	tttgatatcgtcattctgttggag
Y48G8AL.8a	acacaagatcgcggcgagacccat	acacaagatcgcgaagagacccat	ttcgcttgcgcctttaaata	gtgaaccttcgtgatttcgac
Y57G11C.13	tcgatcgcggcgaaacccgtcctcgaaa	tcgatcgcgaagaaacccgtcctcgaaa	aaacccgtcctcgaaactg	tcttgaatattgatgttgaatgag

Table 4.8 – Primers Used for Generation of GFP Constructs

Primers for the three GFP constructs generated for each of the eleven genes. All constructs used the same reverse primer near or overlapping the ATG of the tested gene. This was also the same reverse primer that was used by the BC *C. elegans* Gene Expression Consortium. The "Primer Mutating Motif" differs from the "Primer Including Motif" by two bases; the same mutation was introduced in all cases. The "Primer Excluding Motif" was 12 to 62 bases downstream of the "Primer Including Motif".

4.6 Chapter 4 Figures

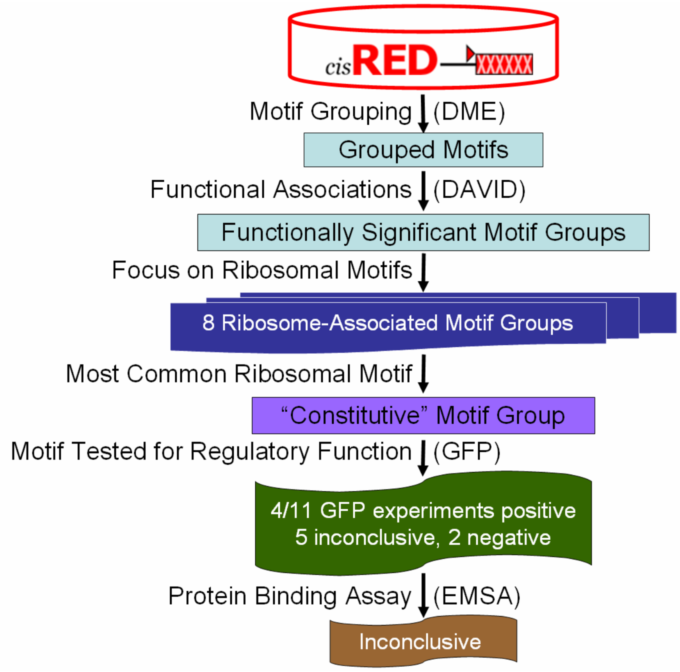

cisRED
XXXXXX
Motif Grouping
(DME)
Grouped Motifs
Functional Associations
(DAVID)
Functionally Significant Motif Groups
Focus on Ribosomal Motifs
8 Ribosome-Associated Motif Groups
Most Common Ribosomal Motif
"Constitutive" Motif Group
Motif Tested for Regulatory Function
(GFP)
4/11 GFP experiments positive
5 inconclusive, 2 negative
Protein Binding Assay
(EMSA)
Inconclusive

Figure 4.1 – Flowchart of Approach

We used the DME algorithm to place the cisRED *C. elegans* conserved element database motifs into groups based on sequence similarity. We then used DAVID to identify motif groups that were associated with genes that also had significant functional similarity. We concentrated our research on eight of the first 20 motif groups that were associated with ribosomal proteins. The largest and most significant of these seemed to be associated with both ribosomal proteins and other constitutively expressed genes (the "constitutive motif"). We tested this motif for regulatory function via a series of GFP constructs using the upstream regions of 11 genes for which we had previous GFP expression data. Four of the 11 genes showed a difference in GFP expression between constructs including the motif and constructs excluding the motif or with a mutated motif. Two of these were tested for protein binding via an electrophoretic mobility shift assay.

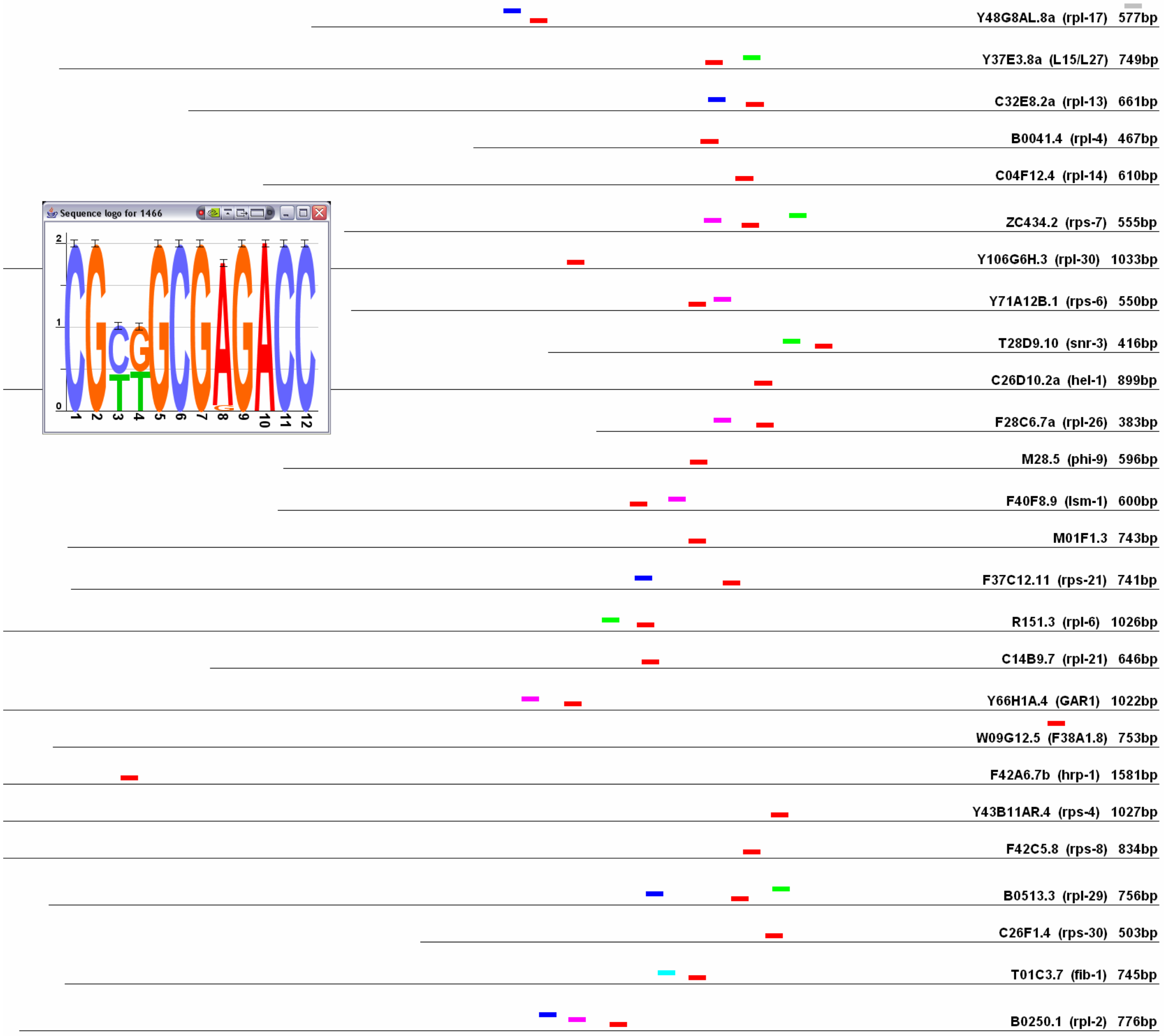


Figure 4.2 – Ribosomal Instances of the Constitutive Motif

The constitutive motif was found upstream of 28 ribosomal transcripts, of which two were on bidirectional promoters. Shown here are the 26 ribosomal upstream regions; instances of motif group 1466 are shown in red. Instances of motif groups 1467, 1471, 1477, 1474, and 1484 are shown in cyan, magenta, grey, blue, and green respectively. The motif logo for all instances of the constitutive motif in these regions is also shown.

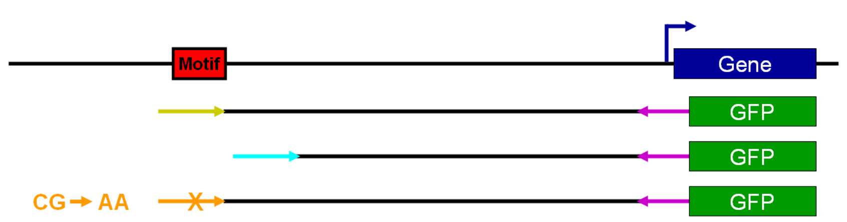


Figure 4.3 – Schematic of GFP Constructs

For each gene with both previous GFP constructs and an instance of the constitutive motif in its upstream region, three constructs were made. The first construct consisted of the gene's upstream region up to and including the motif but no further, the second construct was slightly shorter such that the motif was excluded, and for the third construct, we introduced a mutation in the central CG of the motif via a primer.

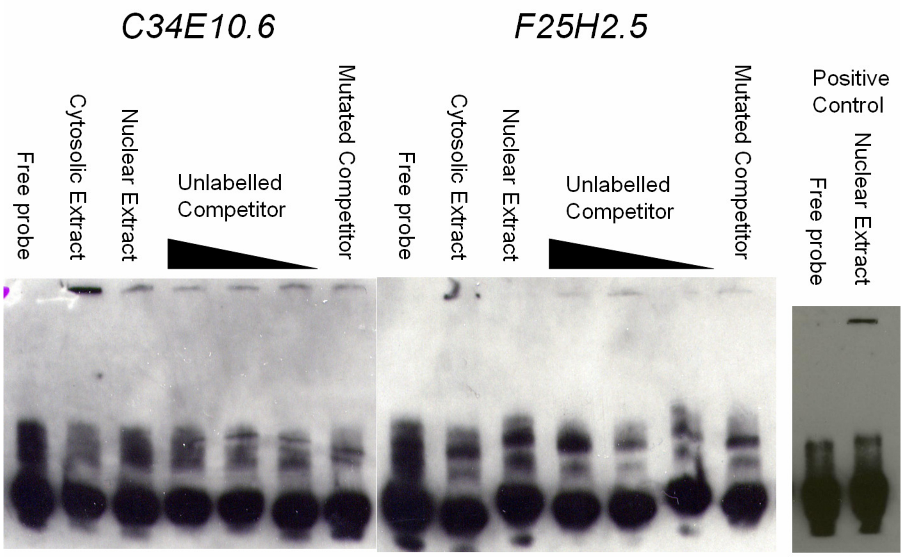


Figure 4.4 – Electrophoretic Mobility Shift Assay

Electrophoretic mobility shift assays were performed on two of the positive results from the GFP experiments. For each assay, we compared the free probe, probe with whole-worm cytosolic extract, and probe with whole-worm nuclear extract. We also performed a competition assay using the labelled probe, nuclear extract, and varying concentrations of unlabelled competitor and mutated competitor. We also performed a positive control using a sequence that had previously been shown to bind protein from the nuclear extract. The lanes in the two EMSAs are as follows (from left to right for each gel):

1. Free biotinylated probe
2. Biotinylated probe + N2 cytosolic extract
3. Biotinylated probe + N2 nuclear extract
4. Biotinylated probe + N2 nuclear extract + 200x molar excess of unlabelled probe
5. Biotinylated probe + N2 nuclear extract + 20x molar excess of unlabelled probe
6. Biotinylated probe + N2 nuclear extract + 2x molar excess of unlabelled probe
7. Biotinylated probe + N2 nuclear extract + 200x molar excess of unlabelled "mutated" probe

4.7 References

1. Sleumer MC, Bilenky M, He A, Robertson G, Thiessen N *et al.* (2009) Caenorhabditis elegans cisRED: a catalogue of conserved genomic elements. Nucleic Acids Res 37: 1323-1334.
2. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36: D107-13.
3. Bryne JC, Valen E, Tang ME, Marstrand T, Winther O *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: 102-106.
4. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: 108-110.
5. Smith AD, Sumazin P, Zhang MQ (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc Natl Acad Sci U S A 102: 1560-1565.
6. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.
7. Hunt-Newbury R, Viveiros R, Johnsen R, Mah A, Anastas D *et al.* (2007) High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. PLoS Biol 5: e237.
8. Eastman C, Horvitz HR, Jin Y (1999) Coordinated transcriptional regulation of the unc-25 glutamic acid decarboxylase and the unc-47 GABA vesicular transporter by the Caenorhabditis elegans UNC-30 homeodomain protein. J Neurosci 19: 6225-6234.
9. Schug J. (2003) Using TESS to predict transcription factor binding sites in DNA sequence. In: Baxevanis AD, editor. Current Protocols in Bioinformatics. J. Wiley and Sons.
10. Obata T, Yanagidani A, Yokoro K, Numoto M, Yamamoto S (1999) Analysis of the consensus binding sequence and the DNA-binding domain of ZF5. Biochem Biophys Res Commun 255: 528-534.
11. Petropoulos I, Auge-Gouillou C, Zakin MM (1991) Characterization of the active part of the human transferrin gene enhancer and purification of two liver nuclear factors interacting with the TGTTTGC motif present in this region. J Biol Chem 266: 24220-24225.
12. Grange T, Roux J, Rigaud G, Pictet R (1991) Cell-type specific activity of two glucocorticoid responsive units of rat tyrosine aminotransferase gene is associated with multiple binding sites for C/EBP and a novel liver-specific nuclear factor. Nucleic Acids Res 19: 131-139.
13. Cardinaux JR, Chapel S, Wahli W (1994) Complex organization of CTF/NF-I, C/EBP, and HNF3 binding sites within the promoter of the liver-specific vitellogenin gene. J Biol Chem 269: 32947-32956.
14. Sekido R, Murai K, Funahashi J, Kamachi Y, Fujisawa-Sehara A *et al.* (1994) The delta-crystallin enhancer-binding protein delta EF1 is a repressor of E2-box-mediated gene activation. Mol Cell Biol 14: 5692-5700.
15. Benson M and Pirrotta V (1988) The Drosophila zeste protein binds cooperatively to sites in many gene regulatory regions: implications for transvection and gene regulation. EMBO J 7: 3907-3915.
16. Suzuki T, Terasaki M, Takemoto-Hori C, Hanada T, Ueda T *et al.* (2001) Proteomic analysis of the mammalian mitochondrial ribosome. Identification of protein components in the 28 S small subunit. J Biol Chem 276: 33181-33195.
17. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B *et al.* (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

18. Fields C (1990) Information content of Caenorhabditis elegans splice site sequences varies with intron length. Nucleic Acids Res 18: 1509-1512.
19. GuhaThakurta D, Palomar L, Stormo GD, Tedesco P, Johnson TE *et al.* (2002) 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: 701-12.
20. Efimenko E, Bubb K, Mak HY, Holzman T, Leroux MR *et al.* (2005) Analysis of xbx genes in C. elegans. Development 132: 1923-34.
21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-9.
22. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J *et al.* (2003) The Protein Information Resource. Nucleic Acids Res 31: 345-347.
23. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M *et al.* (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480-4.
24. Holdeman R, Nehrt S, Strome S (1998) MES-2, a maternal protein essential for viability of the germline in Caenorhabditis elegans, is homologous to a Drosophila Polycomb group protein. Development 125: 2457-2467.

5 Conclusions

5.1 Summary of Major Findings

Chapter 2 contains a description of the bioinformatic methods used in studies of gene regulation in *C. elegans* intestinal cells and sensory neurons. We used the motif discovery algorithms MotifSampler and RSAT to discover a primary motif in the upstream regions of genes that are specifically expressed in intestinal cells. We generated models for the ELT-2 and CHE-1 binding sites using a combination of experimentally validated binding sequences and motif discovery results. We then scanned the upstream regions of *C. elegans* genes to determine the score distribution of sequences that matched the binding models. Finally, we graphed the relative frequency of high-scoring sites in various sets of genes in a cumulative distribution function.

We showed that ELT-2 is the primary transcription factor (TF) responsible for gene regulation in the *C. elegans* intestine, and that intestinally expressed genes are more likely to have a high-scoring ELT-2-like site in their upstream regions than other genes. Similarly, we showed that CHE-1 is the primary but not the only TF responsible for gene regulation in *C. elegans* ASE neurons, and that the CHE-1 binding site is necessary and sufficient for the expression of most ASE-expressed genes.

More generally, we confirmed that coexpressed genes often share transcription factor binding sites (TFBSs) in their upstream regions, and that regulatory elements can be found by motif discovery in the upstream regions of coexpressed genes. We modelled TFBSs with position frequency matrices and found new TFBS candidates by using the PFMs to scan for high-scoring matches.

In Chapter 3, we predicted orthologues in the other seven nematode species using WABA, and identified conserved elements in the orthologous upstream regions using the motif discovery algorithm MotifSampler. The resulting motifs were used to generate the cisRED *C. elegans* database. We validated the motif discovery process using previously found TFBSs from the ORegAnno database and annotated motifs as similar to known TFBSs from the ORegAnno, TRANSFAC, and JASPAR databases.

We found that 17% of *C. elegans* protein-coding transcripts have clearly identifiable orthologues in at least three of the other seven species. The 82% sensitivity showed that most TFBSs are conserved in orthologous upstream regions. Notably, we found that the intergenic regions of highly conserved orthologous genes are highly similar, and that experimentally validated TFBSs were not the most-conserved portions of the upstream regions. Many of the conserved elements were similar to TFBSs from *C. elegans* and other species, while others were identified as coding exons and ncRNAs.

In Chapter 4, we placed cisRED motifs into groups based on sequence similarity, simultaneously identifying sequences and sequence variations that are conserved in the orthologous regions much more often than expected. We identified motif groups that occurred in the upstream regions of genes that also had significant functional associations. Many of the significant functional associations were with ribosomal proteins: We identified seven ribosome-associated motif groups and eight variations of the trans-splice acceptor site that are also associated with ribosomal protein genes. Finally we designed and generated a series of GFP expression construct experiments to test the regulatory function of the largest, most significant ribosome-associated motif group.

We found that many sequences appear numerous times among cisRED motifs, indicating that they are conserved more often than expected. Some groups of genes that share the same conserved sequence in their upstream regions were also similar in terms of function as annotated by the Protein Information Resource [1] and Gene Ontology [2]. Ribosomal genes in particular were often found to have two or three of a set of seven sequences in their upstream regions, about 300 bp upstream of the translation start site (ATG). Ribosomal genes also tended to have one of four trans-splice acceptor site variations about 20bp upstream of the ATG. Four of eleven GFP expression construct experiments showed a dependency on the presence of the constitutive motif for pharyngeal expression. We also tested the motif for protein binding via an electrophoretic mobility shift assay (EMSA). We were unable to determine whether or not the constitutive motif binds a protein *in vivo*.

5.2 Discussion

5.2.1 Revisit Assumptions

Transcriptional control is a primary element of gene regulation. The coexpressed gene projects showed that presence of a specific site in the upstream region is related to mRNA levels in the tissue as measured by SAGE. However, the motif discovery approach is limited and can not be used to elucidate all aspects of gene regulation. During the course of the research described here, it has become clear that chromatin compaction, chromatin organization, and histone modifications play major roles in gene regulation [3]. Because more than 99% of DNA sequences that match a TFBS do not bind a TF *in vivo* [4], an important way in which these epigenetic factors may impact TF-controlled transcription is by influencing which regions of the DNA are available for TF binding and transcription in a particular tissue or time.

Transcription is controlled by TFBSs. In the ASE neuron project, we performed a number of experiments to show that the CHE-1 binding site was both necessary and sufficient for the ASE expression of a number of genes. In the GFP expression experiments in Chapter 4, we showed that presence of the constitutive motif was related to pharyngeal expression. However, we did not show that the constitutive motif was bound by a TF.

The mechanism of the interaction between the TFBS (bound by TFs) and the transcription start site (TSS) is explained by the looping model. Several independent experiments have shown that the mediator complex and the RNA polymerase II complex at the TSS interact directly with the specific TFs bound to sites farther upstream, enabling transcription to be initiated [5-7].

We do not yet understand how fine-grained the control of transcription is by means of this mechanism. *C. elegans* contains 600 to 900 genes for proteins with putative DNA binding domains [8,9], many of which may be involved in transcriptional regulation. However, binding sites are known for fewer than 50 of these TFs. Therefore it is reasonable to focus on TFBSs as a primary mechanism of gene regulation even though it is not the only component involved.

TFBSs are mostly found in the upstream regions of genes in nematodes. Of the 192 *C. elegans* TFBSs in ORegAnno, 159 (83%) were within 1500bp of the ATG, while 20 (10%) were further upstream and 13 (7%) were further downstream. Similarly, a few of the CHE-1 binding sites found by scanning deletion mutagenesis were too far upstream to find with our PFM scan. For example, the CHE-1 binding site responsible for the ASE expression of *gcy-6* was beyond the next-most upstream gene. However, overall it is clear that the immediate upstream region is specifically enriched for TFBSs while the other regions are not.

TFBSs consist of 6 to 14 bp conserved motifs. All examples of TFBSs in *C. elegans* were less than 17 bp wide, with 96% in the range of 6 to 14 bp, so this underlying assumption is still valid. However, 11% of the conserved upstream elements were 20bp or wider, with the result that the TFBSs could not be clearly identified if they were within these wide conserved regions.

Orthologous and coexpressed genes are likely to contain the same TFBSs. We found numerous examples of conservation of TFBSs in the upstream regions of both orthologous and coexpressed genes. The upstream regions of intestinal and ASE-expressed genes were much more likely to contain ELT-2-like and CHE-1-like sites than genes in general. When motif discovery was performed on the upstream regions of the *C. briggsae* and *C. remanei* orthologues of intestine-specific genes, the same motif was found as for *C. elegans*.

5.2.2 Unexpected Findings

The most surprising finding was the degree of conservation in the upstream regions of the orthologous genes. An unstated assumption of the orthologous upstream analysis was that conserved upstream elements would be sparse enough that the TFBSs would be prominent in the conserved portions of the upstream regions. The *Caenorhabditis* species are more evolutionarily distinct from each other (in terms of substitutions per site) than humans are from mice and other mammals [10]. With four or five species to compare, the probability that long stretches of DNA would be conserved by chance is very low, but we found that most upstream regions contained multiple regions of highly conserved sequence. This means that the intergenic regions are not under neutral evolution but are under active conservation to a large degree.

We also found that although TFBSs are conserved, they are not the most-conserved portions of upstream regions. For reasons that remain unclear, sequences of unknown function remain highly conserved while TFBSs are only conserved in the positions important for TF binding. The practical result of these findings is that orthologous conservation is not good evidence for regulatory function, even when five or more nematode genomes are compared.

The most interesting findings were the ribosomal-associated motif groups. We expected to re-discover previously known TFBSs among the motif groups; and in fact we may have. The most significant hexameric motif group was the sequence GATAAG, which looks similar to the ELT-2 binding site. We did not explore most motif groups and do not know at this point how many overlap with known sites or with motif annotations. However, finding so many motif groups that were associated with ribosomal proteins was completely unexpected. To our knowledge, nobody has investigated the regulation of ribosomal genes in any species, so this could be a completely new discovery. Although we were not able to determine exactly what the functions of the ribosomal motif groups were, their p-values were so significant that they could not have occurred by chance. They are clearly worthy of further investigation.

5.3 Future Directions

The results of the intestinal and ASE neuron projects left a number of unanswered questions that could be investigated in the future. For example, we could examine the regulation of genes that were highly expressed in intestine and ASE neurons but did not have high-scoring matches to the PFMs. Initially, we could search for matches further upstream, in introns, and further downstream to confirm which genes are clearly not regulated by ELT-2 and CHE-1. We could then perform motif discovery on the upstream regions of these genes to search for a secondary motif involved in their regulation. We may be able to find a novel site that is the target of one of the other TFs that are present in those tissues. We could also perform a chromatin immunoprecipitation using ELT-2 and CHE-1 to experimentally validate the high-scoring matches that we found. This would improve the binding models of the two TFs and produce a larger set of positive controls for future motif discovery efforts.

An important follow-up to the cisRED database motif annotation result would be experimental validation of the annotated motifs. For example, we could investigate motifs that are both similar to the CHE-1 binding sites and also are associated with neuronal genes. Using GFP expression constructs and chromatin immunoprecipitation, we could determine whether these motifs have regulatory function and bind CHE-1 as predicted. We could also perform a statistical analysis on the annotated motifs to determine which TFBSs tend to co-occur in the upstream regions of the same genes.

As described in the Chapter 3 discussion, there are several unexplored applications of the cisRED database. Investigating the importance of some of the very wide motifs could lead to the discovery of more unannotated exons and ncRNA genes. The deeply conserved motifs could be investigated for conservation in more distant species such as arthropods and vertebrates.

Only 12% of *C. elegans* protein-coding genes have annotated 5’ untranslated regions (UTRs), primarily due to the incertitude of the transcription start site caused by trans-splicing. We used the translation start site to define the gene start in the cisRED project, partly because so few genes have 5’ UTRs and partly because WABA, being optimized only for protein-coding sequence, can only find orthologous sequence starting from the ATG. The result is that many cisRED motifs are in the 5’ UTRs of their respective genes (for those genes that have annotated 5’ UTRs), and this leads us to a number of further questions: Are motifs in 5’ UTRs more or less likely to be annotated as similar to TFBSs compared to motifs in true intergenic regions? Are they more or less likely to be in motif groups? How well conserved are the 5’ UTRs?

The motif grouping results were fairly preliminary and left many unanswered questions. The constitutive motif requires further exploration for us to fully understand its significance. The most obvious experiments to perform would be to use more GFP expression constructs to test the constitutive motif for regulatory function in the upstream regions of ribosomal genes. Because it seemed to co-occur with other motif groups, it would be interesting to test all of the motifs in the upstream region of each ribosomal gene. An alternative GFP expression test would be to concatenate several copies of the constitutive motif (or a combination of several ribosomal motif groups) and attach them to GFP driven by a low-expression core promoter to see whether they are capable of inducing ectopic expression in any tissues.

Further EMSAs are necessary to determine whether any of the ribosomal motif groups bind proteins *in vitro*. In our EMSA experiments, we only used a very short piece of DNA; we may get a more conclusive result if we used a longer piece of DNA containing the constitutive motif together with instances of the other motif groups. If we obtained a positive EMSA result, we could follow up that experiment with a biochemical purification of the binding protein followed by mass spectrometry in order to identify it.

Many of the motif groups that had significant functional associations were unexplored in this research and may provide further insight into gene regulation in *C. elegans*. In particular, motif groups that were significantly associated with homeobox genes, transit peptides, and anatomical structure development would be suitable for further investigation. Additionally, a statistical analysis of co-occurrence could be performed for the motif groups as was suggested for the annotated motifs.

Many interesting developments have materialized in the field of *C. elegans* gene regulation during the course of this research. The Marian Walhout group at the University of Massachusetts Medical School has published a series of papers about *C. elegans* TFs and their binding sites as determined by yeast-1-hybrid experiments [9,11-13]. The purpose of these experiments is to determine which TFs are involved in the regulation of which genes. The findings of the Walhout group could be combined with the results of the cisRED database to determine which uncharacterized *C. elegans* TFs might bind to cisRED motifs.

Recently, a *Caenorhabditis* Genome Analysis Consortium has been formed to generate a comprehensive analysis of the five *Caenorhabditis* genomes. Anticipated topics of investigation include: gene and orthologue prediction, ncRNA gene prediction, repeat detection, whole-genome alignments, protein families, and regulatory sequence analysis. (Erich Schwartz, personal communication). Over twenty research groups are participating in the project and preliminary results will be presented at the 17th International *C. elegans* Meeting at UCLA in June 2009. Similarly, the National Human Genome Research Institute recently approved funding for the modENCODE project, a large-scale analysis of gene regulation for both *C. elegans* and *Drosophila melanogaster* (www.modencode.org). Aspects of this project include transcriptome analysis, chromatin function, histone variants, regulatory elements, and 3' UTRs. It will be interesting to compare the results of both the consortium and the modENCODE project with the results of the similar but much less extensive *C. elegans* cisRED database.

Lastly, this research has several applications to human and other mammalian systems. Mammalian cisRED databases already exist for human, mouse, and rat genomes [14]. Although conserved upstream elements were found for these species, and motifs were annotated for similarity to known sites from TRANSFAC and JASPAR, the motif grouping with DME was not performed on these databases. Given the remarkable success of the *C. elegans* motif grouping procedure, the same techniques for the mammalian cisRED databases should produce interesting results. Also, the mammalian motifs were not examined for similarity to *C. elegans* TFBSs such as CHE-1. We could see if the human genome contains an orthologue of CHE-1, whether it has a similar binding site in the human genome, and whether it is also involved in the regulation of neuronal genes in mammals.

5.4 Philosophy

5.4.1 Innovation

Many of the techniques used in this research were well-established. MotifSampler and RSAT have each been cited more than 200 times and have been used for regulatory element discovery in a wide range of species. The mammalian cisRED database, which was developed by a large team of researchers at the Genome Sciences Centre, was published in 2006 and has since been cited more than 40 times. Motif discovery, position frequency matrices, and motif scanning are common methods to explore the regulation of small sets of coexpressed genes. Comparative genomics is a relatively long-established field, especially for mammalian genomes. However, several aspects of this research were entirely novel.

We have described the first large-scale analysis of multiple nematode genomes. The genomes used in these analyses were made available by several genome sequencing organizations, but we were the first to analyze more than three of them simultaneously. We are also the first to apply mammalian transcription factor binding models to *C. elegans* to see whether our understanding of gene regulation in mammals can be applied to nematodes. We are also the first to use DME secondarily to the results of another motif discovery program, and also to use motif discovery on an input set (the cisRED *C. elegans* motif sequences) that did not originate from coexpressed genes, orthologous genes, chromatin immunoprecipitation sequence, or any other set of sequences that are known to share TFBSs. The cisRED motifs were simply a set of short sequences that are partly or wholly conserved in the orthologous regions of other nematodes, but they were all determined independently and are otherwise unrelated.

It may not have been possible to find the trans-splice acceptor sites and other ribosomal-associated motif groups using any other method. Although ribosomal genes are known to be coexpressed [15], they tend to display a high constant level of expression in all tissues, rather than different levels of expression under different conditions, so the coexpression may not be considered interesting. We did not find any publications regarding the regulation or upstream conservation of ribosomal genes, nor did we find any recent findings regarding trans-splice acceptor site variations.

5.4.2 Impact

Both collaborators went on to publish more papers that built on our findings. ELT-2 was confirmed to be the primary TF for intestinal gene regulation [16]. The Hobert lab continued to investigate the significance of the asymmetry between the ASER and ASEL neurons [17]. The cisRED *C. elegans* database is publicly available and designed to be easy to use. People interested in the regulation of a specific gene and targets of a specific TF can find candidates for further investigation. The impact of the motif grouping results remain to be seen. They may represent a small breakthrough in our understanding of ribosomal genes, gene regulation and/or trans-splicing. Alternatively, because very little related research has been published in these areas, it may take some time before the results connect with other scientists' findings.

5.4.3 Bioinformatics

This research would not be possible without recent advances in DNA sequencing technology, computer algorithms, and computer processing speed. Four of the eight genomes were released while the cisRED database project was already underway and the other four were updated annually throughout the course of the project. With sequencing technology increasing in speed and decreasing in cost each year, research projects in the areas of comparative genomics and gene regulation are going to continue to increase in scope. Some of the algorithms we used were already available at the start of the research (e.g. MotifSampler), but others were published and integrated into the analysis pipeline after preliminary cisRED motif discovery results were already complete (e.g. DME). Computer processing speed and storage space continue to improve. The large-scale orthologue predictions and motif discovery portions of the cisRED analysis pipeline would not have been possible without the GSC parallel computing cluster. Advances in all three of these fields will continue to promote our understanding of DNA and other fields in bioinformatics such as systems biology.

Unlike chemistry and physics, biology is not a science that can be deduced from first principles. Instead, we observe a complicated system already in progress (whether it is on the scale of whole geographical regions or biochemical reactions too small to be seen with the microscope) and try to understand what is happening. At every scale of investigation, every aspect of biology turns out to be far more complicated than we expected or imagined. From

Leeuwenhoek's surprising discovery of bacteria in the 1670s to the more recent failure (so far) of protein structure prediction methods, our current understanding can not prepare us for what we are about to discover. Scientific advances are made by investigating something that has not been investigated before, even if its usefulness or importance is uncertain.

Bioinformatics is our attempt to establish a field of theoretical biology. There are only 20 primary amino acids and four DNA bases; both are easy to digitize. DNA sequence is responsible for the astounding complexity of the biosphere, and yet it has no discernible meaning or information outside of its biological context. In order to understand what a given DNA or amino acid sequence might mean, we have to form a simple model, form testable hypotheses, and only test one element at a time. We inevitably find that the model can be used to explain some initial findings, but when the model is extrapolated, it does not fit every case or even a majority of cases. There are always levels of other factors going on that we are not aware of and may not be aware of for years.

When this research was first proposed, it seemed clear that, based on the few TFBSs that had been discovered, many more TFBSs were waiting to be found if we searched for them. While the research was underway, other scientists were discovering that chromatin organization and histone modifications had a major impact on gene regulation that we were not taking into account. Simultaneously, it was discovered that transcription is not limited to protein-coding genes; numerous noncoding RNA genes influence cellular processes and many genes are transcribed in the antisense direction as well as the sense direction. During the course of our own research, we observed that upstream regions are more highly conserved than expected. It is possible that portions of intergenic regions may be conserved simply because they have few opportunities to mutate, not just because they are physiologically important. At the same time, TFs bind to such a wide variety of sequences under different conditions that their binding sites do not stand out in the highly conserved background. Once again, biology has proven to be complex at a different level than our experiments were designed for. We will need to integrate all of the new discoveries to form a completely different model, and continue our investigations.

5.5 References

1. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J *et al.* (2003) The Protein Information Resource. Nucleic Acids Res 31: 345-347.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-9.
3. Barrera LO and Ren B (2006) The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding. Curr Opin Cell Biol 18: 291-298.
4. Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-87.
5. Casamassimi A and Napoli C (2007) Mediator complexes and eukaryotic transcription regulation: an overview. Biochimie 89: 1439-1446.
6. Higgs DR, Vernimmen D, Hughes J, Gibbons R (2007) Using genomics to study how chromatin influences gene expression. Annu Rev Genomics Hum Genet 8: 299-325.
7. Miele A and Dekker J (2008) Long-range chromosomal interactions and gene regulation. Mol Biosyst 4: 1046-1057.
8. Okkema PG and Krause M (2005) Transcriptional regulation. WormBook 1-40.
9. Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA *et al.* (2005) A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol 6: R110.
10. Kiontke K, Gavin NP, Raynes Y, Roehrig C, Piano F *et al.* (2004) Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. Proc Natl Acad Sci U S A 101: 9003-9008.
11. Barrasa MI, Vaglio P, Cavasino F, Jacotot L, Walhout AJ (2007) EDGEdb: a transcription factor-DNA interaction database for the analysis of C. elegans differential gene expression. BMC Genomics 8: 21.
12. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA *et al.* (2006) A gene-centered C. elegans protein-DNA interaction network. Cell 125: 1193-1205.
13. Vermeirssen V, Barrasa MI, Hidalgo CA, Babon JA, Sequerra R *et al.* (2007) Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network. Genome Res 17: 1061-1071.
14. Robertson G, Bilenky M, Lin K, He A, Yuen W *et al.* (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. Nucleic Acids Res 34: D68-73.
15. Griffith OL, Pleasance ED, Fulton DL, Oveisi M, Ester M *et al.* (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. Genomics 86: 476-488.
16. McGhee JD, Fukushige T, Krause MW, Minnema SE, Goszczynski B *et al.* (2009) ELT-2 is the predominant transcription factor controlling differentiation and function of the C. elegans intestine, from embryo to adult. Dev Biol 327: 551-565.
17. Etchberger JF, Flowers EB, Poole RJ, Bashllari E, Hobert O (2009) Cis-regulatory mechanisms of left/right asymmetric neuron-subtype specification in C. elegans. Development 136: 147-160.