

**Evolutionary dynamics of conserved
non-coding DNA elements: Big bang
or gradual accretion?**

Sujai Kumar



Master of Science
School of Informatics
University of Edinburgh

2007

Abstract

Background Previous studies have found that DNA elements are highly conserved in species from the same lineage, even though they do not code for proteins or RNA. One proposed function of such conserved non-coding elements (CNEs) is that they are *cis*-regulatory sequences for developmental genes which act as an abstraction of genetic regulatory networks, thus allowing new animal body plans to be specified in a modular way. This thesis tests the specific proposal by a previous study that CNEs arose in a big bang in the Precambrian, approximately 600 million years ago.

Results The evolutionary dynamics of CNEs were studied by first identifying the elements, and then examining their levels of identity over time. Pairwise comparative sequence analysis of five contemporary nematode species provided a window into the past because these species diverged at different points of time over the last approximately 700 million years. The number of CNEs and their basic properties for the three most recently diverged species match the results obtained by other researchers, although no clear trend is visible in the change in identity of CNEs with respect to time since divergence. On adding two more species to the analysis, it was found that no such elements could be identified for species pairs with deep divergences.

Conclusions The absence of CNEs for pairwise comparisons of species that diverged earliest indicates that CNEs did not arise in a big bang. CNEs that were found for the three *Caenorhabditis* species that diverged relatively recently (approximately 100 million years ago) seem to be specific only to that clade. However, the big bang hypothesis cannot be conclusively discarded because it is possible that the elements exist, but are short, or have multiple components spread across the genome, and are therefore difficult to detect. Missing CNEs could therefore be a limitation of computational approaches to discovering CNEs, and this study also suggests some ways to overcome those limitations.

Acknowledgements

I am very grateful to Alasdair Anthony and Ann Hedley at the Institute of Evolutionary Biology for getting me started on the mechanics of this project. Other members of the lab group also patiently heard my semi-formed thoughts on conserved non-coding elements, asked penetrating questions, and offered useful advice from time to time.

Many thanks are also due to Douglas Armstrong for providing access to excellent computing resources and for helping with all administrative aspects of the MSc course.

Most importantly, I would like to thank Mark Blaxter whose enthusiasm for life and all living things is contagious. He was the inspiration behind this project and provided much advice, encouragement, and pizza over the course of the summer.

Contents

1	Introduction	1
1.1	Conserved Non-coding Elements (CNEs)	2
1.1.1	Conservation of DNA	2
1.1.2	Non-coding regions of the genome	3
1.1.3	Properties and proposed functionality of CNEs	3
1.2	Hypothesis and approach	5
1.3	Scope	7
1.4	Structure	7
2	Methods and Materials	9
2.1	Obtaining genome sequences	9
2.2	Finding CNEs	11
2.2.1	Finding conserved portions	11
2.2.2	Removing coding regions	14
2.3	Determining CNE similarities	18
3	Results	20
3.1	CNE counts	21
3.1.1	CNEs found using methodology and parameters from Vavouri et al.	21
3.1.2	CNEs found after additional steps to remove coding regions	22
3.1.3	CNEs found for higher sensitivity levels	23
3.2	CNEs shared across all pairs of species	25
3.3	CNEs for <i>C. elegans</i> , <i>C. brigssae</i> , and <i>C. remanei</i>	27
3.4	Aggregate properties of CNEs	30

4 Discussion	34
4.1 Rejection of big bang hypothesis	34
4.2 Limitations of current study, and future work	35
Appendix: Coding regions in GFF files	37
Bibliography	40

List of Figures

1.1	Phylogenetic tree of five nematode species compared in this study (<i>Caenorhabditis</i> divergences from Stein et al., 2003; <i>B. malayi</i> and <i>T. spiralis</i> divergences from Vanfleteren et al., 1994)	5
1.2	Identity vs time plots to verify evolutionary dynamics of CNEs. A minimum 25% identity is expected in all cases (dotted line) because sequences are made up of only four bases: A, T, G, and C. Because background nucleotide concentrations are biased (e.g. lower G-C levels), the minimum level of identity would be higher than 25% (dot-dashed line).	6
2.1	Example fragment of a nucleotide FASTA file	10
2.2	Steps for finding CNEs	12
2.3	(a) Fragment of a megablast results file (column headers: q = query identifier, t = target database identifier, %id = percentage identity, len = length of alignment, mis = number of mismatches in alignment, gap = number of gaps, q_st = starting coordinate of query sequence, q_en = query end, t_st = target start, t_en = target end, e = expect value, bit = bit score) and (b) the output of <i>combine-mbl.pl</i> for that fragment.	13
2.4	Ten sample lines of GFF annotation file for <i>C. briggsae</i> Chromosome I. The highlighted entries depict coding regions. If the coordinates of a megablast result overlapped these coordinates, then it was discarded as a conserved coding region. Eventually, only putative conserved non-coding elements (CNEs) remained. See Appendix for the complete list.	14
2.5	Fragment of coding region file with asterisks used to tag GFF source and feature combinations that specified a coding region.	15

3.1	Characteristics of CNEs found for each pair (y axis), plotted against the time since divergence of the species in that pair (x axis): a) length, b) bit-score, and c) percentage identity. Because several pairs share the same time since divergence (such as <i>C. elegans</i> – <i>C. briggsae</i> , and <i>C. elegans</i> – <i>C. remanei</i> , both 100 MYA), this plot jitters the locations along the x-axis to make it easy to identify the data points for each pair.	31
3.2	CNE percentage identity versus length, visualized as a scatterplot and as a 3D histogram, for <i>C. briggsae</i> – <i>C. remanei</i> , <i>C. elegans</i> – <i>C. briggsae</i> , and <i>C. elegans</i> – <i>C. remanei</i> . . .	33

List of Tables

1.1	Level of conservation of CNEs in different groups of species	4
2.1	Sources for whole genome sequences.	10
3.1	CNEs found for all ten pairs (in alphabetical order) of five nematode species using the method and parameters from Vavouri et al. (2007) for finding CNEs (with megablast parameters -W 30, -e 0.001)	21
3.2	CNEs found after additional checks to determine coding regions (-W 30, -e 0.001)	23
3.3	CNEs found using different sensitivity parameter settings.	24
3.4	Results of clustering CNEs found at different sensitivity levels	26
3.5	Comparisons of mean length, bit-scores, and percentage identity for CNEs shared in two comparisons: <i>C. briggsae</i> – <i>C. remanei</i> (diverged 80 MYA), and <i>C. elegans</i> – <i>C. briggsae</i> (diverged 100 MYA)	29
3.6	Comparisons of mean length, bit-scores, and percentage identity for CNEs shared in two comparisons: <i>C. briggsae</i> – <i>C. remanei</i> (diverged 80 MYA), and <i>C. elegans</i> – <i>C. remanei</i> (diverged 100 MYA)	29

Chapter 1

Introduction

The increasing availability of full genome sequences has led to many comparative studies that have examined the non-protein-coding part of genomes. Over the last decade, several non-coding elements have been found that are completely conserved or conserved with a high degree of identity in species as diverse as *Homo sapiens* and *Fugu rubripes* (the Japanese pufferfish) which last shared a common ancestor approximately 450 million years ago (MYA). This level of conservation indicates that such sequences are functional even though they do not code for proteins or RNA.

The real function of such elements remains an open question. Understanding the evolution of non-coding DNA has the potential to address questions such as how genomes evolved and how they are still evolving. More importantly, it gives us a way to attempt to answer fundamental questions such as how the incredible complexity of life came to be.

This thesis analyses the evolutionary dynamics of conserved non-coding elements (CNEs). It builds on the foundation laid by Vavouri et al. (2007) where they proposed that CNEs are regulatory elements for developmental genes and that it was the "rewiring" of CNEs that led to evolution of the vast diversity of animal body plans. In their study, they compared the genomes of three species from the phylum Nematoda: *Caenorhabditis elegans*, *Caenorhabditis briggsae*, and *Caenorhabditis remanei*. Their analysis is replicated here, and two additional species from the same phylum were added for which full genome sequences have recently become available: *Brugia malayi* and *Trichinella spiralis*. These five species last shared a common ancestor more than 600 MYA and the comparative analysis in this thesis helps answer whether CNEs arose only once (in a "big bang") in the Precambrian as proposed by Vavouri et al., or emerged gradually through evolutionary history. Answering this question would provide us with insights into the process of evolution, may allow us to better understand how animal body plans are specified, and let us speculate whether another explosion in species diversity (of the kind seen 600 MYA) is possible in the future.

To test the big bang hypothesis, the main analytical method employed in this thesis was the level of similarity between such elements for species that diverged at different times. In subsections 1.1 and 1.2, the current understanding of CNEs is reviewed as background material for understanding the hypothesis. The hypothesis and the approach used to test it are presented in detail in subsection 1.3. In the last parts of this introductory chapter, the scope and structure of the remaining chapters of this thesis are presented.

1.1 Conserved Non-coding Elements (CNEs)

Conserved non-coding elements (CNEs) are a recent discovery in several cross-species comparisons. Research groups have not yet decided on a common term for them and each has its own acronym for such sequences:

- CNE - Conserved Non-coding Element (Vavouri et al., 2007, 2006, Woolfe et al., 2005)
- UCR - Ultra-Conserved non-coding Region (Sandelin et al., 2004)
- MCS - Multi-species Conserved Sequence (Margulies et al., 2003)
- CNG - Conserved Non-Genic sequence (Dermitzakis et al., 2005)
- HCE - Highly Conserved Element (Siepel et al., 2005)

This thesis uses the term CNE because it builds on the research and claims made by Vavouri et al. (2007, 2006) and Woolfe et al. (2005), and because the term captures both key aspects of such sequences: that they are conserved across species, and that they are non-coding for proteins or RNA.

1.1.1 Conservation of DNA

Conserved DNA sequences are interesting because they indicate that the sequences are functional. Non-functional sections of the genome undergo mutation and drift apart as species diverge away from each other. Functional sections of the genome remain recognisably similar over long time periods because they code for proteins, code for RNA, are structural, or act as regulatory sites for enhancers, promoters, repressors, and so on. If a section of DNA has such a functional role, it will be under purifying selection and is likely to remain the same or similar over millions of years of mutation pressure. This is the key idea behind all comparative analyses across species, and is a way of identifying functional parts of the genome.

1.1.2 Non-coding regions of the genome

Historically, protein-coding genes were the focus of genome studies (Bird et al., 2006) and a sequence was considered interesting only if it was transcribed as a protein or as RNA. As better experimental and informatics technologies were developed, it was discovered that protein and RNA coding genes only account for small proportions of the whole genome (1.5% to 25% in animal genomes). Comparative genomics studies have identified non-coding regions that appear to be highly conserved and although some parts are now understood to be a complex interacting network of regulatory elements, the functionality of other non-coding parts remains unknown.

1.1.3 Properties and proposed functionality of CNEs

CNEs have been identified for groups of vertebrates and invertebrates separately. Although no sequence identity has been discovered so far between CNEs in vertebrates and CNEs in invertebrates, they share characteristics such as:

- High levels of identity (higher than that of protein-coding genes in most cases), across a wide range of species: Table 1.1 summarizes the level of conservation of CNEs for different groups of species. Although the data are from different sources and do not use the same measures of identity, the figures provide a general idea of how much these sequences are conserved even in the case of species that diverged approximately 450 MYA.
- Clustering around genes: The density of CNEs is higher in gene-rich regions in humans (Bejerano et al., 2004, Sandelin et al., 2004, Woolfe et al., 2005) and nematodes (Vavouri et al., 2007), with several CNEs clustered around each gene.
- Association with developmental genes: Gene association is determined by looking for the transcription start site nearest to each CNE. CNE-associated genes seem to be enriched for regulators of development such as transcription factors and signalling genes (Sandelin et al., 2004, McEwen et al., 2006).

CNEs also exhibit other interesting properties that are not yet understood, such as a spike in AT frequency just inside CNE boundaries (in sharp contrast to flanking regions, Vavouri et al., 2007) and that their AT frequencies are similar (~65%) across species despite the background AT content of each genome being different.

Based on the properties listed above (high identity levels, association with developmental genes, specificity to phyla) and on experiments testing the functionality of

Table 1.1: Level of conservation of CNEs in different groups of species

Species Compared	Number of CNEs and Level of Identity	Last Common Ancestor
Fruit Flies (Glazov et al., 2005): <i>Drosophila melanogaster</i> , <i>Drosophila pseudoobscura</i>	20301 elements; 100% identity; Length > 50bp	25-55 MYA
Mammals (Bejerano et al., 2004): <i>Homo sapiens</i> (Human), <i>Mus musculus</i> (Mouse), <i>Rattus norvegicus</i> (Rat)	256 elements; 100% identity; Length > 200bp	55 MYA
Nematodes (Vavouri et al., 2007): <i>Caenorhabditis elegans</i> , <i>Caenorhabditis briggsae</i> , <i>Caenorhabditis remanei</i>	2084 elements; megablast word seed size 30bp (W30) with e-value threshold 0.001; Average length 69bp	100 MYA
Vertebrates (Woolfe et al., 2005): <i>Homo sapiens</i> (Human), <i>Takifugu rubripes</i> (Pufferfish)	1373 elements; 84% identity; Average length ~200bp	450 MYA

CNEs, the most likely function of CNEs is that they are *cis*-elements that regulate the transcription of a core set of developmental regulatory genes in each species.

Cis-elements are regions of DNA that lie on the same strand as the gene they regulate. Genetic regulatory networks (GRNs) use *cis*-elements extensively to regulate the complex production of proteins with the help of biological controls such as signal transducers, switches, feedback loops, feedforward loops, and combinatorial functions such as “and” and “or” relationships (Andrianantoandro et al., 2006). Development of the animal body plan is controlled by large GRNs and changes in core developmental GRNs can result in new animal body plans (Davidson and Erwin, 2006).

Vavouri et al. (2007) proposed that the initial emergence and subsequent modification of CNEs associated with GRNs was responsible for the evolution of new animal body plans. According to this theory, each animal group has a different set of CNEs because the core GRN for that animal group evolved as a result of the rewiring of CNEs. The vast diversity of body plans first seen in the fossil record of the Cambrian indicates that an evolutionary explosion started in the Precambrian and it is possible that the non-coding elements conserved in modern species are “hard-wired” traces of the changes that took place in core developmental regulatory networks. Because CNEs were highly conserved across species in the same family (e.g., mammals, Dermitzakis et al., 2005) and showed no conservation at all across species from different families (e.g., between humans and nematodes), it is plausible that CNEs are linked in this way to the specification of body plans. The claim that CNEs arose in a big bang in a short period of time around the Precambrian and are responsible for the profusion of animal body plans at that time, is interesting and is testable. If the claim is true, it should be possible to identify the same CNEs in all branches of a phylum. The alternative is that CNEs arise from time to time, and that new CNEs can be seen in every branch of the phylogenetic tree.

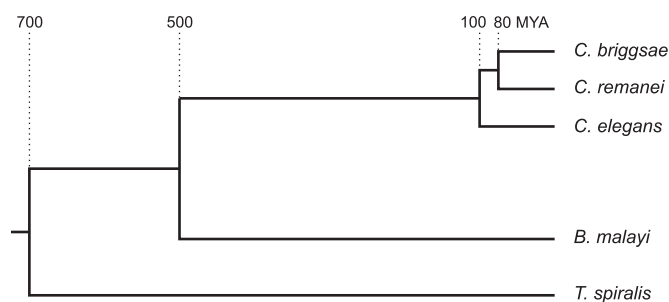


Figure 1.1: Phylogenetic tree of five nematode species compared in this study (*Caenorhabditis* divergences from Stein et al., 2003; *B. malayi* and *T. spiralis* divergences from Vanfleteren et al., 1994)

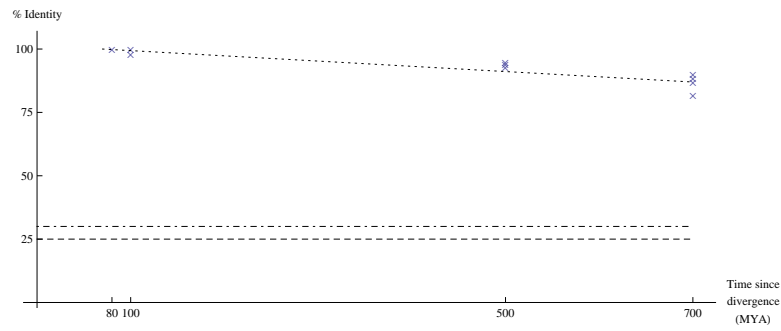
The next subsection frames this hypothesis as a specific question, and outlines the approach taken in this study to determine the evolutionary dynamics of CNEs.

1.2 Hypothesis and approach

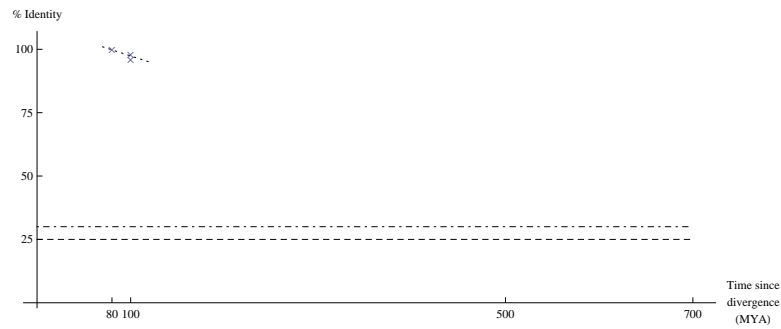
The main goal of this study is to find evidence for or against the idea that CNEs arose in a big bang once, several hundred million years ago. The overall strategy for doing this is to compare genomes from the same family that diverged at different points of time over the last half a billion years, find CNEs in these genomes, and see how their levels of identity have changed over time.

To study the question of how the CNE identities have changed, species from a single phylum are needed because CNEs are not conserved across different phyla. The phylum Nematoda is ideally suited for this purpose because five species within this phylum have recently been sequenced completely, and the approximate dates of divergence for these species span a long time period (from 80 to 700 MYA): *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Brugia malayi*, and *Trichinella spiralis*. The dates in Figure 1.1 are estimates that have an error of ± 20 MYA for the three *Caenorhabditis* species, and the error increases to ± 100 MYA for the branch points of *B. malayi* and *T. spiralis*.

A CNE is found by comparing the genomes of two species, identifying the conserved elements, and removing those elements that are known to code for proteins or RNA. Thus, CNEs are defined for pairs of species for the purposes of this study, and their level of identity for each pair can be determined. Although CNEs are generally conserved with high levels of identity compared to protein-coding sequences, a pair of species that diverged relatively recently (e.g., *C. briggsae* and *C. remanei*, 80 MYA) would be expected to share CNEs with an even higher level of identity than a pair that diverged much earlier (e.g., *C. elegans* and *B. malayi*, 500 MYA).



(a) Supporting big bang model



(b) Supporting gradual accretion model

Figure 1.2: Identity vs time plots to verify evolutionary dynamics of CNEs. A minimum 25% identity is expected in all cases (dotted line) because sequences are made up of only four bases: A, T, G, and C. Because background nucleotide concentrations are biased (e.g. lower G-C levels), the minimum level of identity would be higher than 25% (dot-dashed line).

If CNEs arose in the phylum Nematoda in a big bang, then at least some of the CNEs should be present in all ten pairwise comparisons between the five species, and one would expect the level of similarity for each pair to decrease with the time since divergence of that species. For each CNE, if a plot was created where the x-axis represented time since divergence and the y-axis represented level of similarity (percentage identity), then the plot might look something like Figure 1.2a for each CNE had it arisen in a big bang only once.

On the other hand, if some CNEs are gradually recruited to the genome, then one would expect to see an identity vs time plot as in Figure 1.2b. That is, no corresponding CNEs would exist for pairs that diverged more than a certain number of years ago.

The aim of this thesis is to find CNEs for each of the pairwise comparisons of the five nematode species, and to determine how the identity or similarity of CNEs depends on the time since the pair diverged. Examining the pairwise identities for each CNE would provide evidence for or against the big bang theory of CNE emergence.

It is also possible that some CNEs arose in a big bang during the Precambrian, whereas others arose much later on the evolutionary timeline. Davidson and Erwin (2006)

point out that animal GRNs have levels of hierarchy. Those CNEs associated with the core or kernel GRNs might show big bang features (Fig. 1.2a) because they were responsible for specifying the nematode body plan, whereas those associated with peripheral gene networks might have arisen relatively recently within the individual branches of the phylogenetic tree (Fig. 1.2b).

1.3 Scope

The analysis reported in this thesis draws extensively on existing bioinformatics tools and databases, and several new programs were written to manage the process of finding CNEs and analysing their levels of identity for ten pairs of species. These programs were designed to be very efficient because genome annotation files for well annotated species can be several gigabytes in size and need to be searched rapidly to decide if a particular sequence is coding or non-coding.

Previous research on nematode CNEs had concentrated only on the the three *Caenorhabditis* species: *elegans*, *briggsae*, and *remanei*. The complete genome sequences of *Brugia malayi* and *Trichinella spiralis* have only recently become available and this is the first study to look at conserved non-coding elements in all five species at the same time.

Although past studies have explored other properties of CNEs (such as their AT frequency, gene association, etc.), this study only concentrates on the level of identity or similarity of the CNEs found in each pair of species, because the goal is to look for evidence for or against the big bang hypothesis of CNE emergence.

1.4 Structure

The rest of this thesis is organized as follows. Chapter 2 describes the Methods and Materials used to carry out the study. This part includes details on all the steps used to find CNEs, ranging from descriptions of the data sources to the optimizations carried out to speed up the process of discovering coding regions that had to be eliminated to discover CNEs.

Chapter 3 presents the results, beginning with a summary of the numbers of CNEs found for each pair at different sensitivity levels. The CNEs found are then clustered to determine which CNEs are shared across multiple pairwise comparisons. An analysis follows, describing the trends in identity for several thousand CNEs found in the three *Caenorhabditis* comparisons. The last part of Chapter 3 describes some aggregate properties of the CNEs found.

In the Discussion (Chapter 4), the results are summarized in the context of the original hypothesis. Although the evidence points to a rejection of the hypothesis, the hypothesis cannot be discarded with certainty for several reasons that are presented in detail. The limitations of this study and suggestions for future work complete the thesis.

Chapter 2

Methods and Materials

The following three steps broadly describe the method for testing the hypothesis that CNEs arose in a big bang:

1. Obtain complete genome sequences for the five species being considered
2. Find CNEs
 - (a) Find the conserved portions for each possible pair of species
 - (b) Out of the conserved portions, remove those that overlap known coding regions to identify CNEs
3. Determine CNE similarities for pairs of species that diverged at different points of time

Each step presented its own set of challenges and several choices had to be made for each of the steps above. The challenges, decisions, and the reasons for those decisions are described in the next three subsections.

Several Perl scripts were written for processing the data at each step. These programs are described in this chapter and the source code for the programs is available at <http://www.ylog.org/complex/cne.zip>

2.1 Obtaining genome sequences

Genome sequences for each of the five species were obtained in FASTA format (Pearson and Lipman, 1988) from the sources shown in Table 2.1. FASTA files are a standard

WormBase. The format of these files is described in more detail in Section 2.2.2: Removing coding regions. Although no annotation files for *B. malayi* or *T. spiralis* were publicly available at the time of this study, a FASTA file with all the known coding sequences for *B. malayi* was obtained from Blaxter (2007).

2.2 Finding CNEs

Vavouri et al.'s (2007) methodology was used as the starting point for identifying CNEs. Their procedure for a pair of species was repeated in this study for ten pairs (all possible pairs for five species, as in Table 3.1). The steps and parameters were initially kept identical to verify that the programs for this project were finding CNEs the same way. Subsequently, new parameter sets were tried which are described in Chapter 3. The overall process to find CNEs between two species was to first find the conserved portions (the parts that are recognizably similar) and then to remove those parts that overlap a known coding region. In each of the pairs in the first column of Table 3.1, the first species in the pair had better annotations, and was used as the reference against which coding regions were found and removed.

All the programs ending in `.pl` in Figure 2.2 were developed during the course of this thesis for finding CNEs. These programs are described in the next few subsections (highlighted in italics) along with other publicly available programs that were used at each stage of the process to find CNEs for each pair of species. Additionally, an overall script *pipeline.pl* was written that called these programs with the appropriate parameters for each pair of species.

2.2.1 Finding conserved portions

To identify conserved portions between two species, the program *megablast* (Zhang et al., 2000) was used. *Megablast* takes a query sequence and compares it against a target database to find all the subsequences of the query that have a “hit” (match) against the database. The program uses a heuristic that is much faster than the dynamic programming algorithm for finding sequence alignments (Smith and Waterman, 1981). Although *megablast* is theoretically not guaranteed to find all the alignments between two sequences, in practice it almost always finds the best alignments.

For this thesis, all the parameters for *megablast* were kept at their default values, except the word seed size (`-W`, which specifies the number of contiguous nucleotides that must be identical in an alignment) and the e-value threshold (`-e`, which is a statistical estimate of how often that alignment is likely to occur by chance in a target database

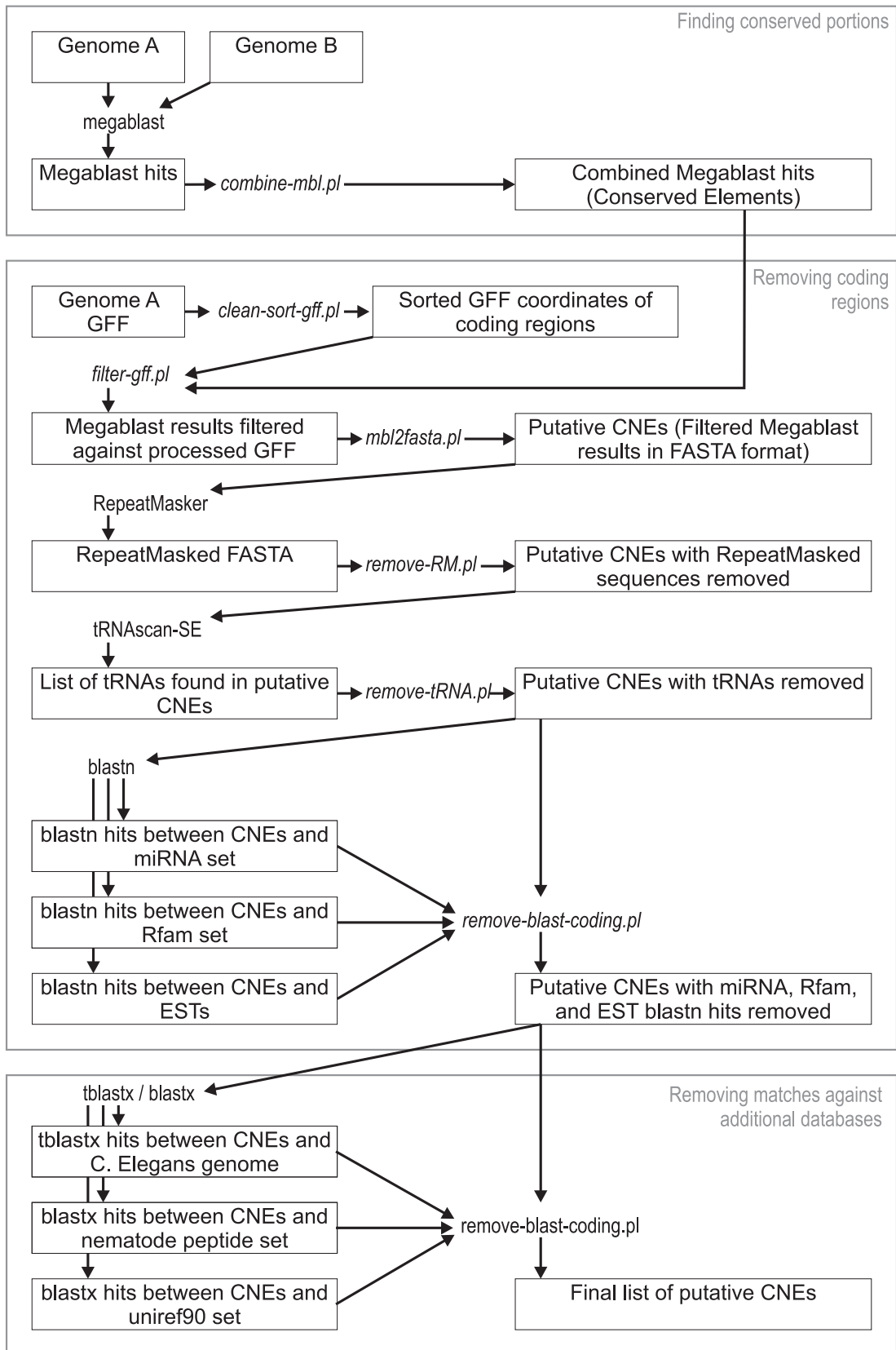


Figure 2.2: Steps for finding CNEs

q	t	%id	len	mis	gap	q_st	q_en	t_st	t_en	e	bit
I	chrI	100.00	74	0	0	3638853	3638926	3011534	3011461	8e-30	147
I	chrIV	97.59	83	1	1	3638845	3638926	3822873	3822955	8e-30	147
I	chrII	100.00	73	0	0	8953163	8953235	3347293	3347365	3e-29	145

(a)

I	chrUn	97.59	-	-	-	3638845	3638926	-	-	-	147
I	chrII	100.00	73	0	0	8953163	8953235	3347293	3347365	3e-29	145

(b)

Figure 2.3: (a) Fragment of a megablast results file (column headers: q = query identifier, t = target database identifier, %id = percentage identity, len = length of alignment, mis = number of mismatches in alignment, gap = number of gaps, q_st = starting coordinate of query sequence, q_en = query end, t_st = target start, t_en = target end, e = expect value, bit = bit score) and (b) the output of *combine-mbl.pl* for that fragment.

of that size). The initial megablast parameters (-W 30 and -e 0.001) were the same as Vavouri et al. (2007), though lower word seed sizes and less stringent e-value thresholds were also tried as reported in the Results chapter. This set of parameters was not very sensitive, and alignments shorter than 30 nucleotides were missed by definition (and some that were longer than 30 were also missed because alignments can have gaps). In comparison, Woolfe et al. (2005) used -W 20 when they found CNEs between human beings and pufferfish. The tabular format was chosen for hits returned by megablast and the better annotated genome was used as the query sequence in all the pairwise comparisons.

Once megablast had run, overlapping hits on the query sequence were combined using *combine-mbl.pl* because it was assumed that two adjacent or overlapping hits represent the same CNE. Figure 2.3 provides an example of how *combine-mbl.pl* works. The query identifier remained the same in the combined megablast result, but the target database identifier was left blank because the hits could have been with different parts of the target database. Target start and end coordinates were also left out for the same reason. The percentage identity and the bit-scores of the combined result were determined by taking the lowest values out of the results that were combined (Dubchak et al., 2000). Finding conserved regions in this way was not symmetric because overlapping regions were combined only for the better annotated genome. However, this was a reasonable simplification because the better annotated genome was the one on the basis of which overlaps with coding regions were determined, as described in the next subsection.

The output of *combine-mbl.pl* was the starting collection of conserved elements (Fig. 2.2). The steps for removing coding regions from this collection are described next.

seqname	source	feature	start	end	score	strand	frameattribute
chrI	curated	intron	9290690	9291216	.	+	.CDS "CBG08484"
chrI	curated	intron	9291336	9292890	.	+	.CDS "CBG08484"
chrI	curated	CDS	223898	229174	.	+	.CDS "CBG11870"
chrI	waba_weak	nucleotide_match	686453	686498	100	.	.Target "Sequence:!" 877470 877515
chrI	waba_coding	nucleotide_match	686499	686522	100	.	.Target "Sequence:!" 877516 877539
chrI	BLAT_briggsae_est	nucleotide_match	1089947	1089958	98.36	+	.Target "Sequence:R03974" 5 17
chrI	BLAT_elegans_est	nucleotide_match	1089948	1089953	73.18	+	.Target "Sequence:yk1518d01.3" 2 8
chrI	curated	exon	2924509	2924669	.	-	.CDS "CBG14900"

Figure 2.4: Ten sample lines of GFF annotation file for *C. briggsae* Chromosome I. The highlighted entries depict coding regions. If the coordinates of a megablast result overlapped these coordinates, then it was discarded as a conserved coding region. Eventually, only putative conserved non-coding elements (CNEs) remained. See Appendix for the complete list.

2.2.2 Removing coding regions

Continuing with the process developed by Vavouri et al, only those conserved portions for each species pair were retained that did not overlap any known coding regions. Identification of the coding regions was a multi-step process that included checking genome annotations, looking for transfer-RNA (tRNA) coding regions, and matching against known expressed sequence tags (ESTs) for that species. Additionally, low-complexity repeats in the genome, and known *elegans* repeats were also marked and removed with the help of the RepeatMasker software package (Smit et al., 1996-2004).

Filtering megablast results against genome annotations

The most important step in deciding if a conserved segment overlapped a coding region was to check it against the genome annotation. Annotations in the General Feature Format (GFF) were downloaded from WormBase (Bieri et al., 2007) for *C. elegans*, *C. briggsae*, and *C. remanei*. These three annotation GFFs were sufficient for checking 9 of the 10 pairwise comparisons, but no GFF was available for *B. malayi* so the *B. malayi*-*T. spiralis* pair was processed differently as described later in this section.

The GFF file fragment in Figure 2.4 lists several fields, but the important ones for this project were the first five: *seqname*, *source*, *feature*, *start*, and *end*. Each line represents a feature at a particular location on the genome. *seqname* was used to identify the chromosome or contig for which the annotation was provided, and the next two fields were used to determine if that annotation was for a coding region or not. The *start* and *end* fields mark the coordinates of that feature on that chromosome or contig.

The Appendix lists the *source* and *feature* combinations found in the three GFF files for *C. elegans*, *C. briggsae*, and *C. remanei*. This list was examined manually and features

<i>source</i>	<i>feature</i>	<i>coding</i>
BLAT_briggsae_est	nucleotide_match	*
BLAT_elegans_est	nucleotide_match	*
curated	CDS	
curated	coding_exon	*
curated	exon	*
curated	intron	
waba_strong	nucleotide_match	

Figure 2.5: Fragment of coding region file with asterisks used to tag GFF source and feature combinations that specified a coding region.

that referred to coding regions were tagged with an asterisk (Blaxter, 2007, Vavouri et al., 2007), and stored in a tab-separated file (Fig. 2.5).

This tab-separated file for identifying exons was then used in program *clean-sort-gff.pl* to pull out all the lines in the GFF file that referred to coding regions. *clean-sort-gff.pl* combined the coordinates of the coding regions (if they overlapped) and only wrote out the start and end coordinates of the combined region to a file that was created for each chromosome or contig referred to in the GFF file. Preprocessing the GFF file in this way into a sorted, non-overlapping set of start and end coordinate pairs for each known coding region was a major optimization. The simplified coding region coordinate file became sufficiently short that it could be loaded into memory, and could be binary searched to see if a megablast result overlapped a coding region. Whereas naive code for checking each megablast result against the entire GFF annotation file took almost 15 hours on a high-end workstation, this optimization sped up the process by a factor of almost 20,000. (e.g., 36,000 megablast results for the *C. elegans*–*C. briggsae* pairwise comparison could be checked against the *C. elegans* GFF with 15 million records in less than 3 seconds).

The *B. malayi*–*T. spiralis* pair of species was tackled differently as no GFF annotation was publicly available for the *B. malayi* genome. The list of conserved elements after the megablast step was converted to a FASTA file (using *mbl2fasta.pl*, described in more detail in the next section) and this FASTA file was “blasted” (i.e., program *blastn*—Altschul et al., 1997—was used to find matches between the two sets of sequences) against a database of known *B. malayi* coding sequences, also in FASTA format. The conserved regions in the *B. malayi*–*T. spiralis* pair that matched the coding sequences for *B. malayi* with an e-value less than 0.0001 were removed, leaving a set of putative CNEs for this pair.

Converting filtered megablast results to FASTA

In the previous step, megablast results for each pair of species were checked against GFF annotations and those that overlapped coding regions were removed. The puta-

tive CNEs were still in megablast output format (specified as a chromosome or contig location, along with the start and stop coordinates). The next set of steps required the putative CNEs to be in FASTA format so that the actual nucleotides could be checked to remove additional coding regions that were missed by the GFF checking step.

Program *mbl2fasta.pl* converted each putative CNE in megablast output format to a FASTA sequence by looking up the appropriate genome sequence file, finding the right chromosome or contig, and pulling out the nucleotides from the start to the stop coordinate. CNEs for the *B. malayi*-*T. spiralis* pair had already been converted into FASTA format in the previous step, so *mbl2fasta.pl* was not run for that pair.

Using RepeatMasker to scan for simple or known *elegans* repeats

RepeatMasker (Smit et al., 1996-2004) is a program that finds interspersed repeats and low-complexity DNA sequences (such as "ATATAT...") and masks these repeats by replacing the repeated nucleotides with a series of Ns. Simple repeats like these are very frequent in the genomes of all species and can lead to uninformative alignments or matches when looking for conserved sequences. RepeatMasker was run on putative CNEs from the previous step at the slowest, most sensitive setting, using *cross_match* (Ewing and Green, 1998) as the comparison engine, and the RepBase library (Jurka et al., 2005) of known repeats for *C. elegans*. The program *remove-RM.pl* was then used to remove all CNEs that contained more than 80% repeats.

RepeatMasker could have been run first for each species before megablast was used to find conserved regions, but it takes a long time to mask repeats in large genomes, so it was more optimal to first run megablast (a very fast algorithm), and then run RepeatMasker only on the conserved regions found.

Using tRNAscan to scan for tRNA coding regions

Once simple repeats had been removed from the putative CNEs, tRNAscan (Lowe and Eddy, 1997) was run to find regions which matched known tRNA coding genes. Some of these regions were removed at the GFF checking stage because some of the better annotations included information on tRNA coding regions. Similar to the previous step, regions identified by tRNAscan as tRNA coding genes were removed using the program *remove-tRNA.pl*.

Removing known Rfam and miRNA regions

Continuing the process of filtering through putative CNEs to remove all known coding regions, the next step was to check the Rfam and micro-RNA (miRNA) databases.

The Rfam database (Griffiths-Jones et al., 2005) has information about families of non-coding RNA and other structural RNA elements, and the miRNA database (Griffiths-Jones, 2004) contains predicted hairpin portions of miRNA transcripts. Putative CNEs were blasted against both these databases (using `blastn` with `-e` set to 0.0001) and any CNEs that showed hits in these two databases were removed using `remove-blast-coding.pl`.

Removing ESTs for poorly annotated species

The final step in Vavouri et al's method for discovering CNEs was to look for matches between the putative CNEs for a pair of species and the Expressed Sequence Tag (EST) databases for those species. An EST is a low-cost, low-quality sequence of nucleotides obtained by sequencing cloned mRNAs. Because they are obtained from mRNAs, a match to an EST is a positive indicator of a coding region, even though it may not be a protein coding gene. EST sequences were downloaded from EBI (Harte et al., 2004) for all the poorly annotated species in this study (i.e. all except *C. elegans*).

As in the previous step, putative CNEs were blasted against the EST sequences for these species, and all sequences that had a match were removed using `remove-blast-coding.pl`.

Finding and removing matches against other databases and genomes

The steps described so far for removing coding regions from a set of conserved elements between two species were proposed by Vavouri et al. (2007). Additionally, Blaxter (2007) suggested blasting (i.e., using programs from NCBI's blast suite of programs) the remaining CNEs against three other sequence sources to be sure that the CNEs obtained were non-coding:

- `blastx` against NemPep3: NemPep3 (Wasmuth and Blaxter, 2006) is an exhaustive database of protein sequences from the phylum Nematoda. `blastx` was used to compare nucleotide sequences against the NemPep3 protein sequence database.
- `blastx` against UniRef90: UniRef90 (Harte et al., 2004) is a non-redundant reference database of all proteins in the UniProt database. Protein sequences with 90% identity are clustered together in UniRef90.
- `tblastx` against *C. elegans* genome (for pairs without *C. elegans*): `tblastx` compares a nucleotide query sequence against a nucleotide target database, after translating all six reading frames of both sets of sequences. Because *C. elegans* had the

best annotated genome, CNEs from pairs that were not checked against the *C. elegans* GFF were checked against the *C. elegans* genome to see if any of the CNEs matched known *elegans* coding regions.

These three searches addressed the same issue: looking for and removing known protein coding regions (especially in nematodes) from the set of putative CNEs. In all three cases, the CNEs with blast hits satisfying an e-value less than 0.0001 were removed using *remove-blast-coding.pl*.

2.3 Determining CNE similarities

The megablast program used for finding conserved regions returns two measures of similarity for each sequence alignment that it finds:

- Percentage identity is the most basic measure of similarity between two sequences and is defined as the percentage of total bases in the alignment that are identical between the two sequences.
- Bit-score is based on the raw score obtained by summing the positive scores for each nucleotide match and the negative scores for each nucleotide mismatch or gap. The raw score is normalized to give the bit-score.

Percentage identity and bit-scores are impossible to derive from the megablast results alone when more than two megablast results are combined (overlapping hits were combined to give a consolidated region according to the procedure by Vavouri et al., 2007). Therefore, based on Dubchak et al. (2000), the lowest percentage identity or lowest bit-score of a set of overlapping hits was taken to represent the overall percentage identity or bit-score of the combined region.

These two measures were used (along with the length of the CNE) to indicate the level of conservation of each pairwise CNE. However, the goal of this thesis was to see if the level of conservation of a CNE changed for different pairs of species. Therefore, a way was needed to establish if two CNEs from different pairwise comparisons represented the same canonical CNE. Blastclust (NCBI, 2007) was used to cluster the CNEs on the basis of their sequence similarities. CNEs belonging to the same cluster had very similar sequences, and, if they came from different pairs of species, then the levels of identity of each CNE could be compared based on the time since divergence of the species in that pair.

For performing all the statistical analyses and creating plots of the levels of identity, Mathematica 6.0 (Wolfram Research Inc, 2007) was used. Mathematica's list and set processing capabilities were especially useful in selecting clusters that contained CNEs from the desired pairs of species.

The results of finding the CNEs, and comparing their similarities across different species pairs based on how long ago the pair diverged, are reported in the next chapter.

Chapter 3

Results

This chapter presents the results of using the pipeline described in Chapter 2 and systematically varying the parameters for finding CNEs. The first part presents the counts of CNEs that were discovered for each pair of species at different levels of sensitivity, and the second part presents CNEs that were shared across multiple pairwise comparisons. One of the main goals of this study was to identify the shared CNEs and see how their level of identity changes depending on how long ago the species they came from had diverged. Although a few CNEs were found for pairs of species that diverged furthest in the past, none of those CNEs are seen in other pairwise comparisons, and so it is impossible to study how those CNEs have changed in identity over time. The third part of the results concentrates on the properties of CNEs for the three species for which the most CNEs were found: *C. elegans*, *C. briggsae*, and *C. remanei*. Unfortunately, these three species diverged from each other relatively recently (100 MYA compared to 500 MYA for *B. malayi* and 700 MYA for *T. spiralis*) so they cannot be used to trace a long evolutionary history of CNEs. The fourth and final part describes some aggregate properties of the CNEs found in each pairwise comparison.

The results are clear that very few or no CNEs were found for pairs of species that diverged earliest, and that none of the thousands of CNEs found for the three species that diverged most recently were shared beyond that group. This is the main finding of this study and is presented in the Discussion (Chapter 4) with more context, along with the implications of such a finding.

Table 3.1: CNEs found for all ten pairs (in alphabetical order) of five nematode species using the method and parameters from Vavouri et al. (2007) for finding CNEs (with megablast parameters -W 30, -e 0.001)

Pair	mega-blast hits	Combine mega-blast hits	Filter against GFF of better annotated species	Remove repeat-masked	Remove tRNA	Remove Rfam, miRNA, and EST matches
<i>B. malayi</i> – <i>T. spiralis</i>	4268	308	34*	15	12	1
<i>C. briggsae</i> – <i>B. malayi</i>	2850	656	395	91	54	21
<i>C. briggsae</i> – <i>C. remanei</i>	23188	14919	7386	6955	6892	6686
<i>C. briggsae</i> – <i>T. spiralis</i>	977	323	173	51	44	22
<i>C. elegans</i> – <i>B. malayi</i>	2784	531	73	14	0	0
<i>C. elegans</i> – <i>C. briggsae</i>	36110	12223	2888	2483	2461	2429
<i>C. elegans</i> – <i>C. remanei</i>	16067	10315	2245	2192	2183	2173
<i>C. elegans</i> – <i>T. spiralis</i>	852	228	19	3	3	0
<i>C. remanei</i> – <i>B. malayi</i>	5258	673	510	127	117	4
<i>C. remanei</i> – <i>T. spiralis</i>	2436	350	263	65	59	7
All pairs	94790	40526	13986	11996	11825	11343

*For *B. malayi*, no annotation GFF was available, so the megablast hits were blasted against a database of known *B. malayi* coding sequences, and matches were removed.

3.1 CNE counts

3.1.1 CNEs found using methodology and parameters from Vavouri et al.

Table 3.1 summarizes the numbers of CNEs found at each stage of the pipeline. The last column lists the number of CNEs found after all the steps proposed by Vavouri et al. (2007). The first step for finding CNEs as described in Chapter 2 is to use megablast to find the conserved regions and this table displays the CNEs found with word seed size 30 (-W 30) and e-value threshold 0.001 (-e 0.001) as described in Vavouri et al.

The sixth row in Table 3.1 (*C. elegans*–*C. briggsae*) acts as a confirmation that the programs written for this project did what they were supposed to do. The number of CNEs found for the pairwise comparison between *C. elegans* and *C. briggsae* corresponds well with Vavouri et al’s findings of 3061 putative CNEs using the same method between the same pair of species. To further verify that the pipeline worked as intended, the set of *elegans* and *briggsae* CNEs found was “blasted” against the set of CNEs published as supplemental online material by Vavouri et al. (i.e., the program blastn was run with stringent settings and all the CNEs matched consistently). The number of putative CNEs for that pair reported here is lower than their finding because of more annotation information available for *C. elegans* since their study, more ESTs for *C. briggsae* in Harte et al. (2004) that were checked against, and updated genome sequences for *C. elegans* and *C. briggsae*.

It is immediately apparent from the last column in Table 3.1 that no CNEs were found at this sensitivity level (-W 30, -e 0.001) for two pairs (*C. elegans*-*B. malayi*, and *C. elegans*-*T. spiralis*), practically none were found for three other pairs (*B. malayi*-*T. spiralis*, *C. remanei*-*B. malayi*, and *C. remanei*-*T. spiralis*), and very few were found for two other pairs (*C. briggsae*-*B. malayi*, and *C. briggsae*-*T. spiralis*).

The only pairs for which thousands of CNEs were found were the pairwise comparisons between the three *Caenorhabditis* species - *elegans*, *briggsae*, and *remanei*. These three comparisons resulted in thousands of CNEs found, more in keeping with the numbers expected on the basis of past studies. The pairs for which no CNEs were found at all using Vavouri et al's methodology and their levels of sensitivity for detecting conserved regions, were the ones between the best annotated species (*C. elegans*) and the species that diverged earliest from the others (*B. malayi* and *T. spiralis*). This finding may indicate that the few CNEs that were found in other pairwise comparisons with *B. malayi* or *T. spiralis* are the result of poorly annotated genomes where some coding regions have not yet been removed as thoroughly as they were for *C. elegans*.

The big-bang hypothesis that this study set out to verify ("CNEs emerged once, several hundred million years ago") has thus been dealt a severe blow, because the very first look at pairwise CNEs seems to indicate that there are no or very few conserved non-coding elements found when comparing nematode species that diverged more than 100 MYA.

This is a strong claim and is examined in more detail in the Discussion chapter. Just to be sure that the megablast program heuristic was not accidentally leaving some conserved regions out, all the CNEs found in pairwise comparisons between the three *Caenorhabditis* species were "blasted" against the *B. malayi* and *T. spiralis* genomes (using program blastn with e-value threshold 0.0001 and word seed size 30) and no hits were returned. The next two parts of this section report the CNEs found at less stringent levels of sensitivity, with some additional steps to ensure that all coding regions were being identified as best as possible.

3.1.2 CNEs found after additional steps to remove coding regions

Three additional steps were performed to identify and remove coding regions in the conserved portions of pairwise comparisons. For details, see Section 2.2.2 on Removing Coding Regions in the chapter on Methods and Materials. Performing the additional steps (removing putative CNEs which had hits at e-value threshold 0.0001 in blastx searches against NemPep3 and UniRef90, and a tblastx search against the coding regions of the well annotated *C. elegans* genome) gives us Table 3.2. The CNE

Table 3.2: CNEs found after additional checks to determine coding regions (-W 30, -e 0.001)

Pair	CNEs found using Vavouri et al's methodology	CNEs after removing blastx/tblastx matches against NemPep3, UniRef90, and <i>elegans</i> genome
<i>B. malayi</i> – <i>T. spiralis</i>	1	1
<i>C. briggsae</i> – <i>B. malayi</i>	21	21
<i>C. briggsae</i> – <i>C. remanei</i>	6686	5288
<i>C. briggsae</i> – <i>T. spiralis</i>	22	21
<i>C. elegans</i> – <i>B. malayi</i>	0	0
<i>C. elegans</i> – <i>C. briggsae</i>	2429	2419
<i>C. elegans</i> – <i>C. remanei</i>	2173	2171
<i>C. elegans</i> – <i>T. spiralis</i>	0	0
<i>C. remanei</i> – <i>B. malayi</i>	4	2
<i>C. remanei</i> – <i>T. spiralis</i>	7	2
All pairs	11343	9925

counts for each pair from Table 3.1 are reproduced to make it easier to compare the effect of the additional steps. The number of CNEs found after the additional steps reduces for pairwise comparisons between the three *Caenorhabditis* species but remains the same or only slightly less for the others.

3.1.3 CNEs found for higher sensitivity levels

Subsequent counts of CNEs found at higher levels of sensitivity in Table 3.3 are only reported for the complete process (i.e., after the additional steps in Section 3.1.2). This table is perhaps the most critical part of this thesis because it demonstrates the difficulty in finding CNEs for some pairs of species, and the conclusions of this thesis rest on this absence of CNEs. At sensitivity levels comparable to previous studies, no CNEs were found for some pairs, but even at higher sensitivity levels (lower word seed sizes, higher e-value thresholds), the CNEs that are thrown up are likely to be false positives on the basis of this table for the reasons presented below.

Keeping the word seed size at 30 and making the e-value threshold less stringent (by increasing it to 0.1) does increase the number of CNEs found for some pairs, but the pairs outside the *Caenorhabditis* triumvirate remain almost the same (Table 3.3). Decreasing the word seed size to 20 with e-value threshold 0.001 almost doubles the CNEs found for each of the *Caenorhabditis* pairs. However, the two pairs that had no conserved non-coding elements (*C. elegans*–*B. malayi* and *C. elegans*–*T. spiralis*) remain without CNEs, and the CNE counts for comparisons between *C. briggsae* and *B. malayi* or *T. spiralis* remain almost the same. At this sensitivity level (word seed size 20 and e-value 0.001), the only pairs for which the number of CNEs increases dramatically are the pairs with poorly annotated genomes. Because coding regions are more easily

Table 3.3: CNEs found using different sensitivity parameter settings.

Pair	-W 30,	-W 30,	-W 20,	-W 20,	-W 12,	-W 12,
	-e 0.001	-e 0.1	-e 0.001	-e 0.1	-e 0.001	-e 0.1
<i>B. malayi</i> – <i>T. spiralis</i>	1	1	385	12313	555	17602
<i>C. briggsae</i> – <i>B. malayi</i>	21	22	23	70	23	82
<i>C. briggsae</i> – <i>C. remanei</i>	5288	5480	10183	12324	12025	15255
<i>C. briggsae</i> – <i>T. spiralis</i>	21	21	22	65	23	82
<i>C. elegans</i> – <i>B. malayi</i>	0	0	0	24	0	31
<i>C. elegans</i> – <i>C. briggsae</i>	2419	2649	4305	6465	4944	7620
<i>C. elegans</i> – <i>C. remanei</i>	2171	2246	4023	4694	4718	5783
<i>C. elegans</i> – <i>T. spiralis</i>	0	0	0	14	0	18
<i>C. remanei</i> – <i>B. malayi</i>	2	2	385	14161	630	20000
<i>C. remanei</i> – <i>T. spiralis</i>	2	2	354	13459	525	19077
All pairs	9925	10423	19680	63589	23443	85550

identified for well-annotated genomes, it is very likely that the jump in CNE counts is a consequence of relatively poor annotation for *C. remanei*, *B. malayi*, and *T. spiralis*.

Increasing the e-value threshold to 0.1, while keeping word seed size at 20 (Table 3.3) again causes a profusion of CNEs in the three most poorly annotated pairwise comparisons: *B. malayi*–*T. spiralis*, *C. remanei*–*B. malayi*, and *C. remanei*–*T. spiralis*. For word seed 12 and e-value 0.1, the number of CNEs for these three pairs has increased by almost four orders of magnitude whereas most of the other pairs show a gradual increase.

The two pairs for which no CNEs were found initially—*C. elegans*–*B. malayi* and *C. elegans*–*T. spiralis*—throw up a few CNEs when the e-value threshold is increased to 0.1, but show no CNEs even at low word seed sizes of 12 and 20 when the e-value is 0.001.

All of these results seem to confirm that when pairs of species are compared that diverged furthest in the past, and there are good annotations available for one of those species (to identify the coding regions), CNEs are remarkably difficult to find. Increasing the sensitivity for finding conserved regions by lowering the word seed size or increasing the e-value threshold certainly returns more CNEs using this pipeline, but most of the CNEs for the pairs with the least annotation are likely to be false positives for the reasons stated above. Considering that previous studies (Vavouri et al., 2007, Bejerano et al., 2004, Glazov et al., 2005, Woolfe et al., 2005) have reported lengthy CNEs (minimum 50 bp) with very high levels of identity, it is reasonable to use the CNEs found with word seed size 30 and e-value 0.001 as specified by Vavouri et al. for all subsequent analyses.

In summary, thousands of CNEs are found for all three pairwise comparisons between the three nematode species that diverged most recently (*C. elegans*, *C. briggsae*, and *C.*

remanei, last common ancestor approximately 100 MYA), but CNEs are not found (or are very rare) for all pairwise comparisons of species that diverged several hundred million years ago.

3.2 CNEs shared across all pairs of species

The goal at the start of this project was to identify those CNEs that are seen in different pairs of species and see if the level of identity for each pair of species changes with how long ago the pair diverged.

The previous subsection has already shown that some pairs of species (*C. elegans*–*B. malayi* and *C. elegans*–*T. spiralis*) don't exhibit any elements that are conserved and non-coding at the same time. This is contrary to the initial expectation that at least some CNEs would be found for all, or almost all pairs, and that it would be possible to study their evolutionary dynamics.

To further verify that there are no CNEs that are similar across the different pairs studied, the complete list of CNEs found was clustered on the basis of sequence similarity. blastclust (NCBI, 2007) was used and set to cluster the CNEs with an 80% similarity cut-off to decide if two CNEs should be treated as being similar enough to belong to the same cluster. This cut-off seemed reasonable because the lowest identity reported in previous studies of CNEs in other species was 84% (Woolfe et al., 2005). The clustering step was repeated for all parameter settings in Table 3.3.

At word seed size 30, and e-value threshold 0.001 (also represented as -W 30, -e 0.001), blastclust took 9925 CNEs and identified 9056 clusters. Out of these, only 687 clusters had more than one member (i.e., 8369 CNEs did not share enough identity with any of the other CNEs to be clustered with them). Of these 687 clusters only 6 clusters included pairs that either had *B. malayi* or *T. spiralis* as one of the species. The remaining 681 clusters only had CNEs found between *C. elegans*, *C. briggsae*, and *C. remanei*.

Each cluster found by blastclust can be thought of as representing one canonical CNE, and the members of the clusters represent the different occurrences of that CNE. Although each of the 6 CNEs found for *B. malayi* or *T. spiralis* is seen more than once (hence the formation of a cluster), the CNEs are always restricted to the same pair. For example, the first cluster with a *B. malayi* CNE in it had 19 such CNEs, but all were found in the comparison between *C. briggsae* and *B. malayi*. Not even one was seen in another pairwise comparison. Similarly, the remaining 5 CNE clusters are also seen in only one pairwise comparison each. To summarize, no *B. malayi* or *T. spiralis* CNEs were seen in more than one pairwise comparison (Table 3.4, last row).

Table 3.4: Results of clustering CNEs found at different sensitivity levels

	-W 30, -e 0.001	-W 30, -e 0.1	-W 20, -e 0.001	-W 20, -e 0.1	-W 12, -e 0.001	-W 12, -e 0.1
Total CNEs found	9925	10423	19680	63589	23443	85550
BLASTCLUST clusters found*	687	726	1351	756	1490	954
Percentage of clusters with CNEs from all three <i>Caenorhabditis</i> comparisons	11%	11%	13%	22%	14%	0%
Clusters with <i>B. malayi</i> or <i>T. spiralis</i> CNEs	6	6	56	550	77	954
<i>B. malayi</i> or <i>T. spiralis</i> CNEs seen in more than one pair	0	0	0	5	1	3

* Only clusters with more than one member are reported.

In comparison, more than half (57%) of the CNEs discovered in the *C. elegans*–*C. briggsae* comparison are also seen in the *C. elegans*–*C. remanei* comparison, and 11% of all the CNE clusters show CNEs from all three *Caenorhabditis* comparisons. This further supports the conclusion that in nematodes CNEs are only really seen for species that diverged within the last 100 million years. For species that diverged before that, a small number of CNEs can be found in pairwise comparisons, but they are not CNEs seen in comparisons with any other species, and therefore it is not possible to make a claim about how the identity of CNEs has changed if we look more than 100 MYA.

This clustering was repeated for the CNEs found using different sensitivity parameters (as in Table 3.3) and the summary of that analysis is shown in Table 3.4.

As in Table 3.3, more sensitive and less stringent settings for finding conserved regions in each pair result in CNEs that are seen in more than one pairwise comparison, but at such low word seed sizes (12), they are much more likely to be false positives. At word seed size 12, e-value threshold 0.1, none of the clusters contained pairs from all the *Caenorhabditis* species. All of the 954 clusters formed at this sensitivity had a CNE from the *B. malayi*–*T. spiralis* comparison. One reason for this could be that the *B. malayi*–*T. spiralis* comparison at that sensitivity yielded the most conserved elements, but it is also the worst annotated pair. Therefore, although several similar sequences have been found using this method, it is very likely that the sequences are coding sequences that have just not been documented till now.

Without CNE clusters that span more than the three *Caenorhabditis* pairwise comparisons, it is not possible to confirm the hypothesis that CNEs arose in a big bang, because there are no CNEs whose identity can be tracked over time for species that diverged very long ago compared to species that diverged in the relatively recent past (as in Fig. 1.2a). The evidence so far strongly suggests that CNEs are a phenomenon of species that diverged 100 MYA or less, or that CNEs from earlier divergences are

undetectable. A more comprehensive summary of this evidence is presented in the Discussion section.

3.3 CNEs for *C. elegans*, *C. briggsae*, and *C. remanei*

Sections 3.1 and 3.2 have consistently shown that CNEs exist for the three *Caenorhabditis* species, and are missing or rarely seen in comparisons with *B. malayi* or *T. spiralis* (species that diverged earliest from the rest). This subsection examines the thousands of CNEs that were found for *C. elegans*, *C. briggsae*, and *C. remanei* in more detail, and tries to see if there is any trend in the level of similarity or identity for the pairs that diverged earlier.

Based on the hypothesis in the introductory chapter, it should be possible to detect some simple trends in how the CNEs have diverged. The fundamental idea behind detecting changes in identity is that if a CNE in one pair of species that diverged X years ago performs the same function as a CNE found in another pair of species that diverged more than X years ago, then the CNEs for the former pair that diverged more recently should show greater identity than the latter pair that diverged further in the past.

Based on Figure 1.1, the following times since divergence can be assigned to each pair:

- *C. elegans*–*C. briggsae*: 100 MYA
- *C. elegans*–*C. remanei*: 100 MYA
- *C. briggsae*–*C. remanei*: 80 MYA

To look at the *Caenorhabditis* CNEs in more detail, Vavouri et al's (2007) method was modified slightly. After the initial megablast search to find conserved regions between a pair of species, the megablast hits were not combined to merge overlapping sequences. Leaving this step out made it possible to identify orthologous CNEs in the subsequent clustering step. However, the number of putative CNEs found also increased because some conserved regions were not combined any more. At word seed size 30 and e-value threshold 0.001, 11,089 CNEs were found for all three *Caenorhabditis* comparisons (compared to 9934 in Table 3.2). When the CNEs were clustered (as in section 3.2) using blastclust with the minimum similarity threshold set to 80%, 862 clusters were found. Assuming each cluster represents one canonical CNE, the CNEs in that cluster are the instances of that canonical CNE seen in different pairwise comparisons. Of the 862 clusters or canonical CNEs found for these three pairwise

comparisons, only 78 (9%) of these clusters had CNEs from all three pairwise comparisons thus indicating that these 78 CNEs were 'universal' for the three *Caenorhabditis* species.

A pairwise sampling statistical method was used to test if any of the correlates of identity (percentage identity, length, or bit-score) consistently varied for species that diverged 80 MYA (*C. briggsae*–*C. remanei*) as compared to species that diverged 100 MYA (*C. elegans*–*C. briggsae*, and *C. elegans*–*C. remanei*). Each cluster of CNEs had CNEs from different pairwise comparisons, and some of the larger clusters had multiple CNEs from the same pair. The following procedure was used to determine how, for example, the percentage identity of a CNE varied when plotted against time since divergence:

1. The set of all clusters with any one *C. elegans*–*C. briggsae* CNE in them was labelled CeCb. Similarly, the set of all clusters with any one *C. elegans*–*C. remanei* CNE was labelled CeCr, and the set with any one *C. briggsae*–*C. remanei* pair was labelled CbCr. These were not mutually exclusive sets, and contained many overlaps.
2. To compare the mean percentage identity of CNEs found in the *briggsae*–*remanei* comparison (diverged 80 MYA) with the mean percentage identity of CNEs found in the *elegans*–*briggsae* comparison (diverged 100 MYA), the intersection of sets CbCr and CeCb was taken. This resulted in a set of 306 clusters, each of which had at least one *briggsae*–*remanei* CNE and one *elegans*–*briggsae* CNE.
3. The next step was to compare the percentage identity of the CbCr CNEs with the percentage identity of the CeCb CNEs. To ensure that only orthologous CNEs from each of the sets were compared, a systematic pairing method was developed:
 - (a) Each CNE cluster in the intersection of CbCr and CeCb was picked one at a time.
 - (b) For each cluster, the set of CbCr CNEs in it and the set of CeCb CNEs in it were extracted as two subsets. Some CNE clusters in the intersection of CbCr and CeCb had only two CNEs, so each CNE would be in a subset on its own.
 - (c) Each CNE from the first subset was paired with each CNE from the second subset. If, for example, the CbCr subset for that CNE cluster had two CNEs and the CeCb subset had three CNEs, then the total number of pairs would be six.
 - (d) Only those pairs were retained where the CNEs for the species in common (e.g., in the case of CbCr and CeCb, *C. briggsae* is the species in common)

came from the same chromosome and shared the same coordinates. This step increased the likelihood that the CNEs being compared were orthologous.

- (e) The percentage identity (or other measure of identity) of the second CNE in the pair was subtracted from the first CNE in the pair.
- (f) Repeating steps (a) to (g) for each pair in each cluster, gives a set of data points representing the difference in the percentage identity.
- (g) If there was no systematic difference in the percentage identities of CNEs from CbCr and the CNEs from CeCb, then the data points representing the difference in identities calculated in step (e) would have mean 0. However, if the mean was statistically significantly greater than 0, that would imply that the CbCr CNEs were systematically more identical than CeCb CNEs.

4. Step 3 was repeated for other correlates of identity such as length and bit-score.

Table 3.5: Comparisons of mean length, bit-scores, and percentage identity for CNEs shared in two comparisons: *C. briggsae*–*C. remanei* (diverged 80 MYA), and *C. elegans*–*C. briggsae* (diverged 100 MYA)

	Length	Bit-scores	% identity
Mean difference between <i>C. briggsae</i> – <i>C. remanei</i> , and <i>C. elegans</i> – <i>C. briggsae</i> CNEs	+0.6	+9.3	+1.13
Statistically significant (t-test)	No (p-value $\simeq 0.05$)	Yes (p-value $\simeq 0$)	Yes (p-value $\simeq 0$)

Table 3.6: Comparisons of mean length, bit-scores, and percentage identity for CNEs shared in two comparisons: *C. briggsae*–*C. remanei* (diverged 80 MYA), and *C. elegans*–*C. remanei* (diverged 100 MYA)

	Length	Bit-scores	% identity
Mean difference between <i>C. briggsae</i> – <i>C. remanei</i> , and <i>C. elegans</i> – <i>C. remanei</i> CNEs	+1.3	-3.3	-0.4
Statistically significant (t-test)	No (p-value $\simeq 0.06$)	No (p-value $\simeq 0.24$)	No (p-value $\simeq 0.17$)

On performing the procedure as described, the mean difference in percentage identity for *C. briggsae*–*C. remanei* CNEs compared to *C. elegans*–*C. briggsae* CNEs was found to be +1.13 (which was statistically significant with a two-sided p-value of 10^{-16} , effectively 0). In other words, if the same CNE was seen in two pairwise comparisons that diverged at different points of time, the level of identity for the pair that diverged only 80 MYA (*briggsae*–*remanei*) was, on average, greater than the level of identity for the pair that diverged 100 MYA (*elegans*–*briggsae*). On repeating this procedure for CNE bit-scores (the scores assigned by the megablast program that takes into account the length, level of identity and the other parameters that were used to do the megablast), *briggsae*–*remanei* CNEs have higher bit-scores (9.3 greater, on average) than the corresponding *elegans*–*briggsae* CNEs. The length of the CNEs is also found to be 0.6

greater for the pair that diverged only 80 MYA as compared to the pair that diverged 100 MYA, but this difference was not statistically significant. Table 3.5 summarizes these differences.

Unfortunately, although the method is innovative and helps to understand the trend for pairs of species that diverged at different points of time, the trend in identities is not consistent. Table 3.6 should have shown the same trend as Table 3.5 if CNEs diverge gradually as more time passes, but here some of the comparisons are negative. Although the mean length of CNEs is higher, the mean bit-score is lower, and mean percentage difference is lower for the species that diverged more recently than species that diverged further in the past. Additionally, none of the comparisons in Table 3.6 are statistically significant at $\alpha = 0.01$ using a t-test (or using Wilcoxon's Signed Rank test).

The three pairwise comparisons for *Caenorhabditis* species only provided two time-points against which some measures of identity could be tested. If more CNEs had been found in all the pairwise comparisons, then the kind of analysis presented here for the three pairwise comparisons could have been carried out for the other pairs of species as well.

To summarize, although several hundred CNEs exist which are seen in all three *Caenorhabditis* pairwise comparisons, no consistent trends in identity can be detected for these CNEs based on the time since the pairs diverged.

3.4 Aggregate properties of CNEs

Although no CNEs were identified that were shared across all (or even most) pairwise comparisons, and it is therefore not possible to see how CNEs for each pair change when compared with pairs that diverged at different times, it is possible to provide some basic statistics for the CNEs that were found for each pair.

Vavouri et al. (2007) have already done a comprehensive analysis of the CNEs that they found by comparing the *C. elegans* genome to the *C. briggsae* genome and the CNEs found in that comparison were further filtered against the *C. remanei* genome to give a set of what they called *wCNEs* (short for worm-CNEs).

Figure 3.1 shows the means (along with errors) of lengths, bit-scores, and percentage identities for all CNEs found in the 10 pair-wise comparisons performed in this study (at $-W 30$, $-e 0.001$). As reported in Section 3.1, two pairs had no CNEs (*C. elegans*-*B. malayi*, and *C. elegans*-*T. spiralis*), therefore only 8 data points are shown. One pair had only one CNE (*B. malayi*-*T. spiralis*) and it shows up as a lone dot on each of the plots.

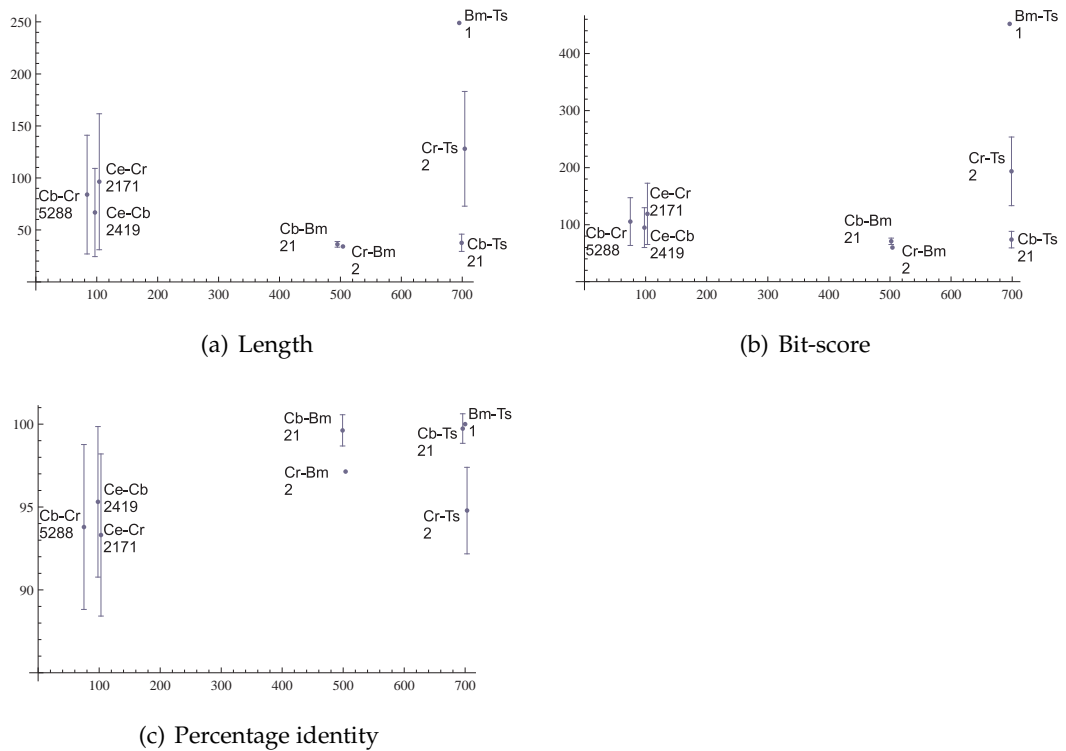


Figure 3.1: Characteristics of CNEs found for each pair (y axis), plotted against the time since divergence of the species in that pair (x axis): a) length, b) bit-score, and c) percentage identity. Because several pairs share the same time since divergence (such as *C. elegans*–*C. briggsae*, and *C. elegans*–*C. remanei*, both 100 MYA), this plot jitters the locations along the x-axis to make it easy to identify the data points for each pair.

No clear trends are visible in any of the measures for each pair when plotted against the time since divergence of that pair.

The *C. elegans*–*C. briggsae* data points in each of the plots in Figure 3.1 (mean length: 67 bp, mean percentage identity: 95.3%) correspond well with the aggregate properties reported by Vavouri et al. (2007)—mean length: 69 bp, mean percentage identity: 96%—further verifying that their method for one comparison has been successfully reproduced and replicated for nine other pairwise comparisons. If more CNEs had been consistently found for other pairwise comparisons, it would have been worthwhile to report and compare other aggregate properties of wCNEs (such as AT frequency inside and outside the CNE boundaries, association with genes, etc), but given that the other pairs resulted in fewer than thirty CNEs compared to the thousands found in the three *Caenorhabditis* comparisons, it is most likely that these CNEs in other pairs are false positives rather than genuine functional conserved non-coding elements.

Finally, Figure 3.2 provides the distributions of the CNEs found for the three *Caenorhabditis* comparisons based on their percentage identity and length. The scatterplots and

3D histograms both show the same data, although the histograms (with Log frequencies) make it obvious that a large proportion of CNEs are 100% identical (*C. briggsae*–*C. remanei*: 20%, *C. elegans*–*C. briggsae*: 30%, *C. elegans*–*C. remanei*: 16%).

All three pairwise comparisons show the same pattern of CNE distributions. The fingerprint-like arcs to the lower right of the scatterplots are not significant. They are an artifact of measuring identity in percentages when the sequences are only approximately 30 nucleotides long. Each mismatch or gap causes a jump in percentage identity and those cause systematic spaces to appear in the scatterplots for low CNE lengths.

Both the scatterplots and the histograms show how the choice of megablast parameters affects the ability to find conserved regions with low sequence length and low percentage identity.

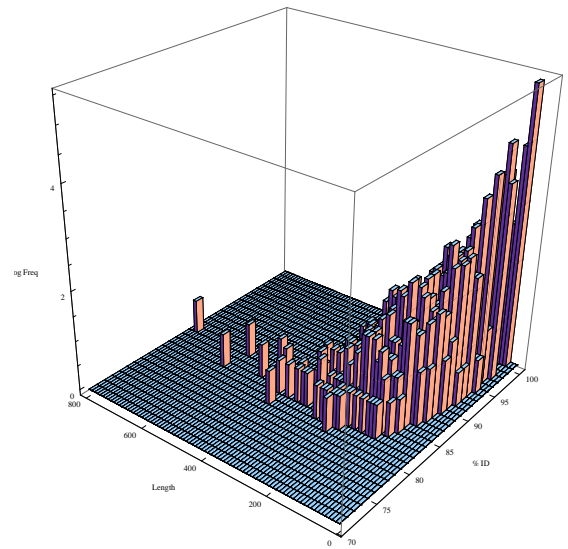
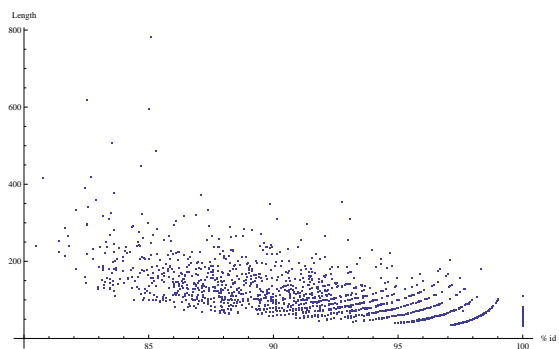
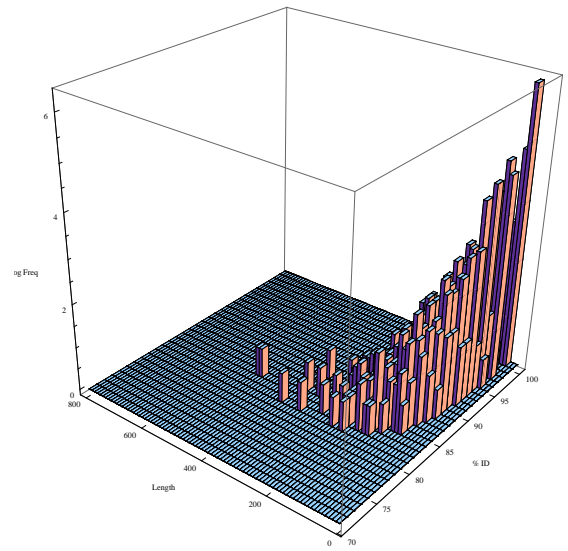
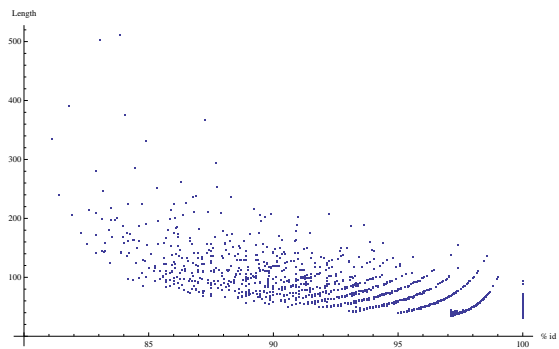
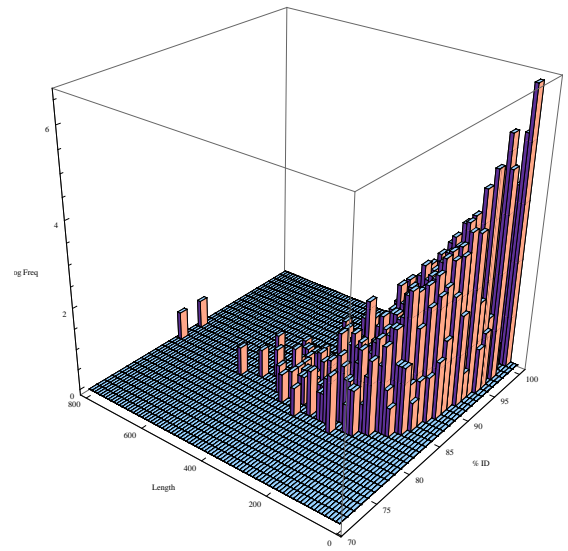
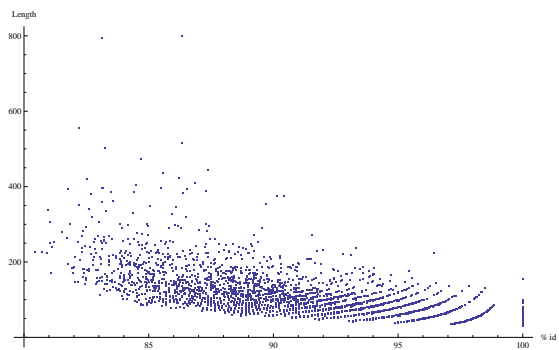


Figure 3.2: CNE percentage identity versus length, visualized as a scatterplot and as a 3D histogram, for *C. briggsae*–*C. remanei*, *C. elegans*–*C. briggsae*, and *C. elegans*–*C. remanei*

Chapter 4

Discussion

4.1 Rejection of big bang hypothesis

Before starting the analysis, the expectation had been that CNEs would be found for all ten pairwise comparisons of the five species studied. To support the big bang hypothesis, several of the same CNEs would have had to be seen across all the pairs. For such CNEs (that were seen in all or almost all pairwise comparisons), a decrease in the level of identity versus time since divergence would have indicated that at least some CNEs arose at the origins of the phylum Nematoda and were carried forward through evolutionary history (Fig. 1.2a). However, not even one putative CNE was found that was seen in all pair-wise comparisons. Therefore no evidence was found of a big bang where the CNEs arose in a common ancestor and subsequently diverged at a low or zero rate of mutation.

It was also surprising that no CNEs at all were found for some of the pairs of species that diverged 500 and 700 MYA when using the sensitivity parameters used by previous studies. Higher sensitivity parameters were also tried for detecting conservation (lower word seed sizes and higher e-value thresholds) and those parameters resulted in many thousands of CNEs for pairs that had initially shown only 1 or 2 CNEs. However, the pairs that showed this dramatic increase in CNEs were also the ones with the lowest levels of annotation, indicating that the increase in CNEs was a result of not being able to identify coding regions. This is further supported by the fact that when *C. elegans* (with the best annotated genome) was compared with the species that it diverged earliest from (*C. elegans*–*B. malayi*: 500 MYA, and *C. elegans*–*T. spiralis*: 700 MYA), no CNEs were found at e-value 0.001 even for word seed sizes as low as 12.

Thus, based on the hypothesis and approach at the start of this study, the results show no evidence for a big bang of CNEs. On the other hand, the absence of CNEs at these

levels of divergence is not conclusive proof that a big bang did not occur, because it is possible that all CNEs arose initially in some nematode ancestor and were subsequently lost or were transformed into other *cis*-regulatory elements in a way that is difficult to detect. Previous studies in vertebrates have shown that non-coding elements remain highly conserved, but it is possible that functional non-coding elements in invertebrates evolve faster than their counterparts in vertebrates. Recent work on rapid divergence in *cis*-regulatory regions (Borneman et al., 2007) also suggests that some non-coding elements might be diverging faster than orthologous genes in the same species.

Additionally, if CNEs are *cis*-regulatory elements, it is also possible that the functional part of a CNE is shorter than the sensitivity parameters tried, or that there are many short (< 20 bp) functional parts that are conserved interspersed with segments that are free to mutate, making CNEs difficult to detect.

In the detailed analysis of the three *Caenorhabditis* pairwise comparisons for which many CNEs were consistently found, no clear trend in identity versus time since divergence was visible. The CNEs in the pair that diverged most recently (*C. briggsae*–*C. remanei*, 80 MYA) were more similar than the *C. elegans*–*C. briggsae* CNEs (100 MYA), but less similar than *C. elegans*–*C. remanei* CNEs (100 MYA), although the latter comparison was not statistically significant. Given that the time since divergence of these species is not certain and that the error of the estimates of 80 and 100 MYA is ± 20 MYA, the most likely reason for a lack of trend is that both pairs of species diverged approximately at the same time, and therefore the level of identity of the CNEs is more or less the same in both pairs.

4.2 Limitations of current study, and future work

This study explores whether CNEs arose in a big bang, using the level of identity of CNEs in pairwise species comparisons as the analytical tool. Multiple indicators or methods might have been able to provide a better, more conclusive answer to the hypothesis. As more experimental studies are conducted on CNEs and more information becomes available about their structure and how they work, other methods could be devised to test this hypothesis.

Given the constraint that CNE identity is the only tool currently available to test the hypothesis, other limitations in this study along with suggestions for overcoming them in future work are listed below.

1. Assumptions in determining conservation: The initial step for finding conserved regions assumes that CNEs consist of identical contiguous nucleotides of a cer-

tain minimum length (the word seed size parameter). This assumption is reasonable (it is the way all genome conservation studies are presently conducted) but may not be valid for the kinds of *cis*-regulatory elements that define animal body plans. When more information is available about CNE structure, future studies should attempt to find conserved regions based on methods that don't rely on contiguous runs of identical elements, such as Hidden Markov Models.

2. Limited annotations for all species other than *C. elegans*: The most likely reason for the profusion of CNEs for some pairs of species at higher levels of sensitivity was that the species in those pairs were poorly annotated. With more extensive and accurate information about coding regions in each species, CNEs can be identified much more confidently, and it is possible that future work will reject the thousands of CNEs found at less stringent e-value levels in this study for some pairs of species.
3. Limitations of megablast as a tool for discovering conservation: Visualizing the distributions of the CNEs (by percentage identity and length) as scatterplots and histograms made it clear how the choice of the initial megablast parameters (for detecting conserved regions) affects the ability to find CNEs. With *-W 30* and *-e 0.001*, no short CNEs (< 40bp) were found with an identity lower than 95%, and this is a limitation of the method for discovering CNEs which may in reality be shorter and have lower levels of percentage identity.

In summary, based on what is currently known about CNEs, this study did not find any evidence that CNEs arose in a big bang at the start of the phylum Nematoda. Future developments such as better genome annotations or better ways of discovering conservation in genomes would reject some of the CNEs discovered till now or allow more CNEs to be nominated respectively, and the hypothesis could be more conclusively supported or rejected in those cases.

Appendix: Coding regions in GFF files

Tab-separated file *coding.txt* specifying which GFF *source-feature* combinations represented a coding region (Section 2.2.2)

<i>source</i>	<i>feature</i>	<i>coding</i>
.	Sequence	
Allele	complex_change_in_nucleotide_sequence	
Allele	deletion	
Allele	insertion	
Allele	sequence_variant	
Allele	SNP	
Allele	substitution	
Allele	transposable_element_insertion_site	
binding_site	miRanda	
binding_site	misc_feature	
binding_site	PicTar	
BLAT_BAC_END	nucleotide_match	
BLAT_briggsae_est	nucleotide_match	*
BLAT_elegans_est	nucleotide_match	*
BLAT_elegans_mrna	nucleotide_match	*
BLAT_elegans_ost	nucleotide_match	*
BLAT_EST_BEST	EST_match	*
BLAT_EST_OTHER	EST_match	*
BLAT_mRNA_BEST	cDNA_match	*
BLAT_mRNA_OTHER	cDNA_match	*
BLAT_ncRNA_BEST	nucleotide_match	*
BLAT_ncRNA_OTHER	nucleotide_match	*
BLAT_NEMATODE	translated_nucleotide_match	*
BLAT_NEMBASE	translated_nucleotide_match	*
BLAT_OST_BEST	expressed_sequence_match	*
BLAT_OST_OTHER	expressed_sequence_match	*
BLAT_TC1_BEST	nucleotide_match	
BLAT_TC1_OTHER	nucleotide_match	
BLAT_WASHU	translated_nucleotide_match	*
cDNA_for_RNAi	experimental_result_region	
Coding_transcript	coding_exon	*
Coding_transcript	exon	*
Coding_transcript	five_prime_UTR	*
Coding_transcript	intron	
Coding_transcript	protein_coding_primary_transcript	
Coding_transcript	three_prime_UTR	*
Coding_transcript	Transcript	

contig	Sequence	
curated	CDS	
curated	coding_exon	*
curated	exon	*
curated	gene	
curated	intron	
Expr_pattern	reagent	
Expr_profile	experimental_result_region	
Genbank	region	
gene	gene	
gene	processed_transcript	
Genefinder	CDS	
Genefinder	coding_exon	*
Genefinder	exon	*
Genefinder	intron	
GeneMarkHMM	CDS	
GeneMarkHMM	coding_exon	*
GeneMarkHMM	exon	*
GenePair_STS	PCR_product	
Genomic_canonical	region	
history	CDS	
history	coding_exon	*
history	exon	*
history	intron	
history	Pseudogene	
history	Transcript	
inverted	inverted_repeat	
landmark	gene	
Link	region	
mass_spec_genome	translated_nucleotide_match	
miRNA	exon	*
miRNA	miRNA_primary_transcript	
Mos_insertion_allele	transposable_element_insertion_site	
mSplicer_orf	CDS	
mSplicer_orf	coding_exon	*
mSplicer_orf	exon	*
mSplicer_transcript	CDS	
mSplicer_transcript	coding_exon	*
mSplicer_transcript	exon	*
ncRNA	exon	*
ncRNA	intron	
ncRNA	ncRNA_primary_transcript	
ncRNA	RNAz	
ncRNA	Transcript	
Non_coding_transcript	exon	*
Non_coding_transcript	intron	
Non_coding_transcript	nc_primary_transcript	
Oligo_set	reagent	
operon	operon	
Orfeome	PCR_product	
polyA_signal_sequence	polyA_signal_sequence	
polyA_site	polyA_site	
predicted	polymorphism	
predicted	rflp_polymorphism	
Pseudogene	exon	*
Pseudogene	intron	
Pseudogene	Pseudogene	
RepeatMasker	repeat_region	
RNAi_primary	RNAi_reagent	
RNAi_secondary	RNAi_reagent	
rRNA	exon	*

rRNA	rRNA_primary_transcript	
SAGE_tag	SAGE_tag	
SAGE_tag_genomic_unique	SAGE_tag	
SAGE_tag_most_three_prime	SAGE_tag	
SAGE_tag_unambiguously_mapped	SAGE_tag	
scRNA	exon	*
scRNA	scRNA_primary_transcript	
SL1	SL1_acceptor_site	
SL2	SL2_acceptor_site	
snlRNA	exon	*
snlRNA	Transcript	
snoRNA	exon	*
snoRNA	snoRNA_primary_transcript	
snRNA	exon	*
snRNA	snRNA_primary_transcript	
tandem	tandem_repeat	
TEC_RED	nucleotide_match	
Transposon	exon	*
Transposon	intron	
Transposon	transposable_element	
Transposon_CDS	coding_exon	*
Transposon_CDS	exon	*
Transposon_CDS	intron	
Transposon_CDS	transposable_element	
tRNA	exon	*
tRNA	tRNA_primary_transcript	
twinscan	CDS	
twinscan	coding_exon	*
twinscan	exon	*
twinscan	intron	
Vancouver_fosmid	region	
waba_coding	nucleotide_match	
waba_strong	nucleotide_match	
waba_weak	nucleotide_match	
WU_MERGED	CDS	
WU_MERGED	coding_exon	*
WU_MERGED	intron	
wublastx	protein_match	*

Bibliography

- S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.
- E. Andrianantoandro, S. Basu, D. K. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology*, 2:2006.0028, 2006.
- G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Hausler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, May 2004.
- T. Bieri, D. Blasiar, P. Ozersky, I. Antoshechkin, C. Bastiani, P. Canaran, J. Chan, N. Chen, W. J. Chen, P. Davis, T. J. Fiedler, L. Girard, M. Han, T. W. Harris, R. Kishore, R. Lee, S. McKay, H.-M. Muller, C. Nakamura, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E. M. Schwarz, W. Spooner, M. A. Tuli, K. V. Auken, D. Wang, X. Wang, G. Williams, R. Durbin, L. D. Stein, P. W. Sternberg, and J. Spieth. WormBase: new content and better access. *Nucleic Acids Research*, 35:D506–510, 2007.
- C. P. Bird, B. E. Stranger, and E. T. Dermitzakis. Functional variation and evolution of non-coding dna. *Current Opinion in Genetics & Development*, 16(6):559–564, December 2006.
- M. L. Blaxter. Personal communication, 2007.
- A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder. Divergence of transcription factor binding sites across related yeast species. *Science*, 317(5839):815–819, August 2007.
- E. H. Davidson and D. H. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, February 2006.
- E. T. Dermitzakis, A. Reymond, and S. E. Antonarakis. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nature Reviews Genetics*, 6(2):151–157, February 2005.

- I. Dubchak, M. Brudno, G. G. Loots, L. Pachter, C. Mayor, E. M. Rubin, and K. A. Frazer. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Research*, 10(9):1304–1306, September 2000.
- E. Ghedin, et al. Draft Genome of Filarial Nematode Parasite *Brugia Malayi*. *Science*, In press, 2007.
- B. Ewing and P. Green. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8:186–194, 1998.
- E. A. A. Glazov, M. Pheasant, E. McGraw, G. Bejerano, and J. S. S. Mattick. Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mrna splicing. *Genome Research*, May 2005.
- S. Griffiths-Jones. The microRNA Registry. *Nucleic Acids Research*, 32:D109–D111, 2004.
- S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33:D121–D124, 2005.
- GSC. Washington Univ in St Louis Genome Sequencing Center: *C. remanei*, 2007a. <http://genome.wustl.edu/genome.cgi?GENOME=Caenorhabditis%20remanei>.
- GSC. Washington Univ in St Louis Genome Sequencing Center: *T. spiralis*, 2007b. <http://genome.wustl.edu/genome.cgi?GENOME=Trichinella%20spiralis>.
- N. Harte, V. Silventoinen, E. Quevillon, S. Robinson, K. Kallio, X. Fustero, P. Patel, P. Jokinen, and R. Lopez. Public web-based services from the European Bioinformatics Institute. *Nucleic Acids Research*, 32:W3–9, 2004.
- J. Jurka, V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, 110:462–467, 2005.
- T. Lowe and S. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25:955–964, 1997.
- E. H. Margulies, M. Blanchette, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518, December 2003.
- G. K. K. McEwen, A. Woolfe, D. Goode, T. Vavouri, H. Callaway, and G. Elgar. Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Research*, March 2006.
- NCBI. BLASTCLUST - BLAST score-based single-linkage clustering, 2007. <http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastclust.html>.

- W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *PNAS*, 85(8):2444–2448, April 1988.
- A. Sandelin, P. Bailey, S. Bruce, P. G. Engstrom, J. M. Klos, W. W. Wasserman, J. Ericson, and B. Lenhard. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):5–99, 2004.
- A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, August 2005.
- A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996-2004. <http://www.repeatmasker.org>.
- T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.
- L. D. Stein, Z. Bao, D. Blasiar, T. Blumenthal, M. R. Brent, N. Chen, A. Chinwalla, L. Clarke, C. Clee, A. Coghlan, A. Coulson, P. D’eustachio, D. H. Fitch, L. A. Fulton, R. E. Fulton, S. Griffiths-Jones, T. W. Harris, L. W. Hillier, R. Kamath, P. E. Kuwabara, E. R. Mardis, M. A. Marra, T. L. Miner, P. Minx, J. C. Mullikin, R. W. Plumb, J. Rogers, J. E. Schein, M. Sohrmann, J. Spieth, J. E. Stajich, C. Wei, D. Willey, R. K. Wilson, R. Durbin, and R. H. Waterston. The genome sequence of *caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biology*, 1(2):e45+, November 2003.
- The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, 282(5396):2012–2018, December 1998.
- J. R. Vanfleteren, Y. Van De Peer, M. L. Blaxter, S. A. Tweedie, C. Trotman, L. Lu, M.-L. Van Hauwaert, and L. Moens. Molecular genealogy of some nematode taxa as based on cytochrome c and globin amino acid sequences. *Molecular Phylogenetics and Evolution*, 3(2):92–101, June 1994.
- T. Vavouri, G. K. Mcewen, A. Woolfe, W. R. Gilks, and G. Elgar. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends in Genetics*, 22(1):5–10, January 2006.
- T. Vavouri, K. Walter, W. R. Gilks, B. Lehner, and G. Elgar. Parallel evolution of conserved noncoding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biology*, 8:R15+, February 2007.
- J. Wasmuth and M. L. Blaxter. On the origins of genic novelty in the phylum Nematoda. 2006.

Wolfram Research Inc. Mathematica Edition: Version 6.0, Wolfram Research Inc., 2007.

A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. Edwards, J. E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, 3(1), January 2005.

Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7(1-2):203–214, 2000.