Part II

Studies in Evolution

Chapter 5

Wormdup: a Database of DNA Duplications in *Caenorhabditis elegans*

5.1 Chapter introduction

The evolution of new genes by duplication is a key component of molecular evolution. Of fundamental interest are the mechanisms by which genes are duplicated and the scale on which these duplications take place. There are many examples of searches for large-scale block duplications involving diverse genes (see e.g. [SKR89, WS97, Hug98]), perhaps the most striking of which was Wolfe and Shields' publication of evidence for a tetraploid duplication of the entire yeast genome [WS97]. Numerous examples of local tandem duplications of single genes, giving rise to two or sometimes more daughter genes, are also present in the literature (see e.g. [Sid96, BTR98, FBT+91, Eis98]) and indeed the abundance of this type of duplication is evident from a cursory inspection of the annotation of published genomic sequence.

From a neutralist argument one might expect that gene duplications represent special cases against a background of continuous turnover - involving duplication and reciprocal deletion - of non-coding as well as coding DNA. There are three main processes by which it is recognised that DNA duplication can occur in eukaryotes: (i) polyploidy, whereby an organism acquires a duplicate copy either of a single chromosome or its entire genome [Ohn70, WS97, BB98]; (ii) copying of host DNA during the process of transposon integration [Jur98] or excision repair [MKW91]; and (iii) unequal crossing-over between chromosomes during meiotic recombination [LG91]. The last of these - unequal crossing-over - may be triggered by numerous causes; experimental evidence suggests it can happen quickly where there is an existing tandem gene duplication [BTR98] and it can also be triggered by multiple, adjacent copies of a repetitive element [FBT+91].

Our understanding of the dynamics of gene creation and the relative importances of the different ways DNA can be copied is far from perfect, although population genetic models for these processes have been explored [TK98, Oht91]. With the increasing amount of genomic sequencing the paucity of data is likely

to be replaced by the technical difficulties of gross analyses as the main obstacle to better understanding.

With these issues in mind, a database - "Wormdup" - oriented specifically towards researchers interested in studying aspects of genomic duplications has been constructed. Wormdup contains co-ordinates and age estimates of unique, single duplications of non-coding DNA in the recently sequenced [CSC98] genome of the nematode *Caenhorabditis elegans*. The focus is on non-coding DNA so that general features of duplications may be studied in the absence of gene-specific selective pressures. The score cutoffs used in the creation of Wormdup were chosen so that no duplications large enough to contain a gene should be missed. Various pre-calculated filters on the data are offered, including (amongst others) raw BLAST matches, gapped matches constructed using dynamic programming, duplications involving genes and large repeat families. In addition a suite of tools has been developed to facilitate the construction of more complex custom filters.

Section 5.2 of this chapter describes the structure of Wormdup, including the tools and algorithms that were used in its construction and may be used to query it. In Section 5.3, the Wormdup data are used to calculate various parameters of molecular evolution for *C.elegans*. These include the duplication size and separation distributions, the fixation rate of duplications and the subsequent rates of divergence by stochastic accumulation of substitutions and deletions. Section 5.4 investigates the number of Wormdup entries found in conformations suggestive of repetitive-element-mediated duplication. In Section 5.5, the molecular evolutionary parameters for non-coding duplications are compared to the parameters for coding duplications. It is found that the apparent fixation rate of gene duplications is higher than the rate for non-coding duplications of the same size. The implications of this discrepancy are discussed. In Section 5.6, the results of the evolutionary parameter-fitting are summarised and discussed.

5.2 Methods

This section describes the techniques used in the construction of Wormdup. The section begins with an overview of the nature of the algorithms required for a project of this kind, and proceeds to detail the process of construction of the core units of Wormdup.

The starting point for the entire analysis was the 72Mb of finished *C.elegans* DNA sequence available in May 1998.

A schematic view of the main stages in the construction of Wormdup is shown in Figure 5.1. The Wormdup data files and many of the tools and protocols are available from the following URL:

`http://www.sanger.ac.uk/Users/ihh/Wormdup/`

5.2.1 Overview of methods

Many of the common tasks involved in gross analyses of sequence features can be reduced to a series of manipulations on sets (or ordered sets) of sequence co-ordinates, where a set of co-ordinates for these purposes is defined as a *(name,startpoint,endpoint)* tuple (henceforth NSE). An example of an NSE tuple is the location of the gene AH6.2, which spans residues 5054 to 6308 of *C.elegans* cosmid AH6: the appropriate NSE is *(AH6,5054,6308)* in cosmid co-ordinates and *(CHROMOSOME_II,9624958,9626212)* in chromosome co-ordinates (since cosmid AH6 starts at base number 9619904 on chromosome II, according to the map used for this work).

A useful if basic format for representing lists of NSEs is GFF, the Gene Finding Format, developed in collaboration between the Sanger Centre, the University of California at Santa Cruz and other participants [GFF]. Each line of a GFF flat-file describes a single NSE with some additional information (such as, to continue the above example, the orientation of the gene AH6.2). This extra information is irrelevant to many of the algorithms (though this is by no means a hard-and-fast rule, with scoring information being the most frequent

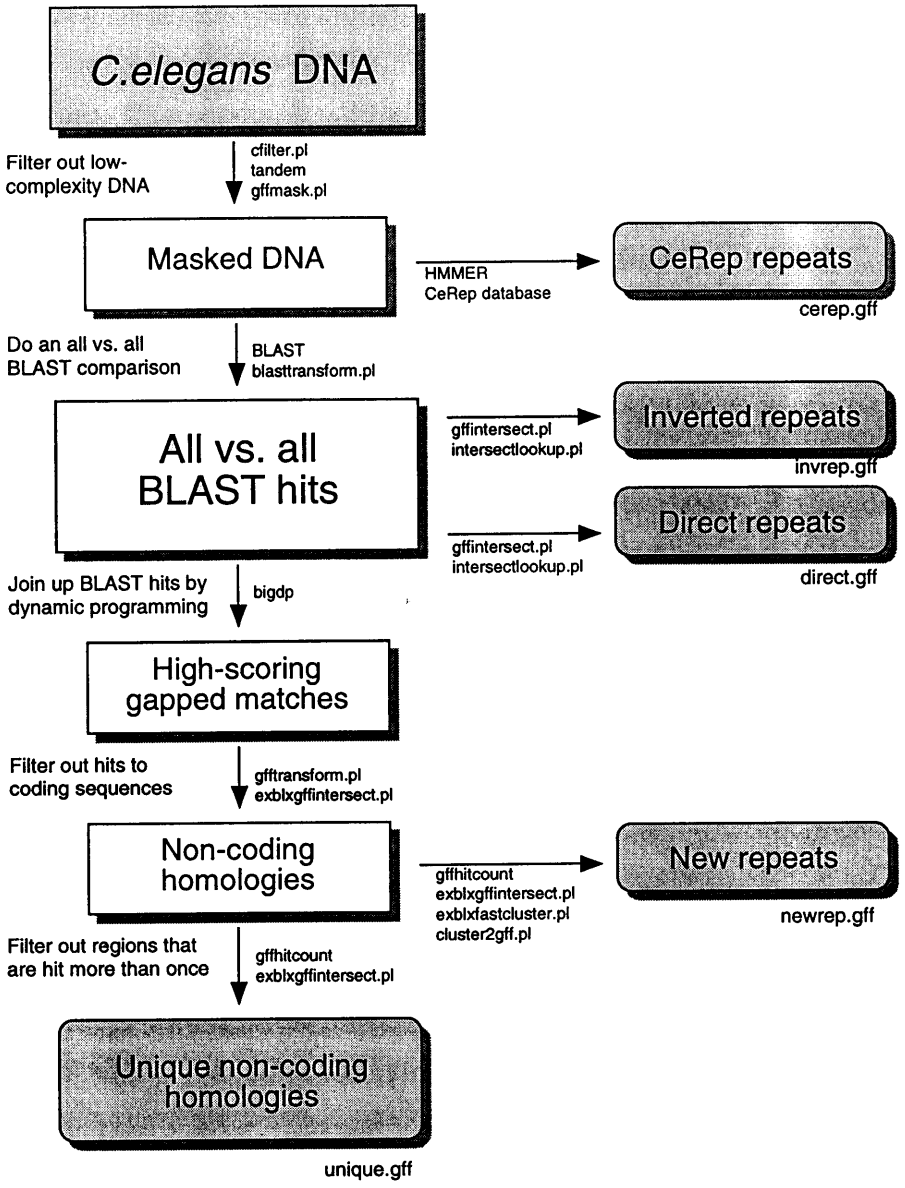

Figure 5.1: Schematic view of the construction of Wormdup. Names of key scripts and programs are shown to the right of/beneath arrows linking intermediate stages. Not all script names are shown, but all relevant scripts are described in Appendix A and available from the Wormdup website. The filenames with ".gff" suffices beneath shaded boxes refer to data files on the website.

exception). Single NSEs are inadequate to represent certain kinds of information (for example, homologies); a GFF-pair protocol exists for this purpose, though EXBLX (an output format of the BLAST post-processor `MSPcrunch` [SD94]) is in some ways preferable as a format for representing NSE-pairs as it is both more compact and more symmetrical. In any case conversion tools between these various formats were developed early in the analysis.

Apart from format conversion, the most common elementary operations that can be performed on (ordered) sets of NSEs include: (i) set intersection, (ii) set exclusion, (iii) set union, (iv) filtration, (v) sorting, (vi) merging (of sorted lists), (vii) transformation (of co-ordinate systems) and (viii) dereferencing (access to the described sequence). With a suitably flexible definition of NSE similarity, these operations form a basis for more sophisticated algorithms like clustering and tiling. Pointers to a comprehensive set of tools and links for manipulating NSEs in GFF and EXBLX format, including the GFFTools programs that were developed for this project, are maintained on the GFF website [GFF].

GFF is a relatively new format at the time of writing, and in many cases the tools described here were the first available for this format. In more cases, they were the fastest, being designed with respect to the consideration that reading millions of NSEs and NSE-pairs into memory at once is not practical.

Unless explicitly referenced, the tools described in the following sections were all developed principally for the Wormdup project. A full description of the GFFTools package may be found in Section A.4 of Appendix A.

Both GFF and EXBLX were found to be adequate formats for the present project; though it is the author's opinion that the single most pertinent feature of both formats, at least for working at the shell level, is the correspondence of a single line to a single feature.

5.2.2 Filtering low-complexity regions

The most commonly used low-complexity filter for use with DNA sequence is the `dust` program; however, the filtering heuristic used by `dust` is somewhat *ad hoc* [Tat]. A slightly more analytically supported method is the sliding-window entropy filter used by the `seg` program [Woo94], but this is designed for protein sequences. For the low-complexity masking of the worm DNA, a parameterisable sliding-window low-entropy filter `cfilter.pl` was specially written in Perl (see Appendix A).

Using the `cfilter.pl` program and the `tandem` program from the GCG package [But98], low-entropy regions (12-mer windows whose single-base compositional entropy did not exceed 0.5 bits) and microsatellites were identified, recorded in GFF format and masked from all subsequent analyses.

5.2.3 Preliminary scan for repetitive elements

A preliminary screen for hits to the CeRep database of repetitive sequence profiles was performed using the HMMER1.7 program. Local inverted and direct repeats (those missed by the `tandem` program) were also searched for by BLASTing each cosmid against itself. The lists of repeats were reduced by looking at the self-intersection of the list and taking the closest sequence-pair of every intersecting set, using the `gffintersect.pl` and `intersectlookup.pl` programs described in Appendix A.

The `tandem` search yielded ~13000 tandem repeat regions; the average length of the tandemly repeated regions was 38 bases and on average there were 9 copies of this region. The number of inverted repeats was greater (~71000); this is probably because the `tandem` program attempts to join up multi-copy repeats whereas the inverted repeats are single copies.

The results of the screen for the CeRep elements are summarised in Table 5.1. A more thorough search for *mariner*-like transposable elements was also performed and is described in detail in Chapter 7.

Repeat type	Copy number in 72Mb	Expected copy number in 100Mb
CeRep10	2944	4088
CeRep11	551	765
CeRep12	2271	3154
CeRep13	1128	1566
CeRep14	1089	1512
CeRep15	713	990
CeRep17	635	881
CeRep18	597	829
CeRep19	2001	2779
CeRep20	504	700
CeRep21	469	651
CeRep22	316	438
CeRep23	1870	2597
CeRep24	2535	3520
CeRep28	540	750
CeRep29	545	756
CeRep30	102	141
CeRep31	145	201
CeRep32	719	998
CeRep33	109	151
CeRep34	2008	2788
CeRep35	1134	1575
CeRep36	782	1086
CeRep37	475	659
CeRep38	1017	1412
CeRep39	93	129
CeRep40	128	177
CeRep41	311	431
CeRep42	760	1055
CeRep43	3737	5190

Table 5.1: Copy numbers of CeRep elements in *C.elegans*.

5.2.4 Finding duplicated blocks

An all versus all ungapped BLAST comparison of the finished *C.elegans* DNA formed the basis for nearly all the rest of the Wormdup database. The search was performed with the version of the program designed for nucleotide-nucleotide comparisons, `blastn`, using the default scoring parameters (+5 for a match, -4 for a mismatch; this ratio corresponds to a Jukes-Cantor substitution distance $kt \simeq 0.16$ with a score-to-likelihood ratio of $S/L \simeq 5.2$). The score threshold for reporting hits was 120; low-complexity and tandem regions (but not CeRep, inverted or direct repeats) were masked out. The BLAST results were converted to EXBLX format by `MSPcrunch` [SD94] then transformed into chromosomal co-ordinates by the `blasttransform.pl` program described in Section A.4 of Appendix A. The data were sorted by chromosome-pair and redundancies and self-hits due to overlaps between cosmids were trimmed using the `exblxsort.pl` and `exblxtidy.pl` programs (also in Appendix A).

The ungapped BLAST hits were joined together by dynamic programming, using a program called `bigdp` that implements a modified version of the Waterman-Eggert algorithm [WE87] requiring $O(n)$ space, with a simple optimisation heuristic that reduces the expected compute time from $O(n^2m^2)$ to $O(n^2m)$ (where n and m are the query sequence lengths). The `bigdp` algorithm is described - with a worst-case scenario - and compared to other large-scale sequence comparison methods in Section A.5 of Appendix A.

The dynamic programming used a gap open penalty of 6 and a gap extend penalty of 0.8; using the score-to-likelihood ratio stated above, this corresponds very roughly to a gap frequency of 0.3 gaps per residue per strand and a geometrically distributed gap length with mean 6 residues. The cutoff score for reporting hits was 600, corresponding to a run of 120 perfect matches using BLAST. This high cutoff will exclude many small duplications (which are expected to be more numerous than large duplications) but it should pick up duplications large enough to potentially encode genes or exons, which are of

primary interest. Hits scoring higher than the cutoff are referred to below as "high-scoring duplications".

5.2.5 Excluding genes and repetitive elements

The complete set of high-scoring duplications includes a large number of sequences with multiple hits. Some of these are matches between homologous genes in a multi-gene family and many more are matches between highly repetitive elements. These matches were excluded from the data set. Gene duplications are treated separately in Section 5.2.6 below and repetitive elements are addressed in Chapter 7.

Duplications involving genes were first identified from the *C.elegans* annotation and excluded. Next, repetitive sequence loci were identified using the `gffhitcount` program (described in Appendix A) which counts the number of times each base on a chromosome is covered by a high-scoring duplication. The distribution of base hit counts is shown in Figure 5.2. In total 48.7Mb (94.7%) bases of non-coding DNA were not hit by any high-scoring duplications at all; 625kb (1.2%) were hit once and 2.10Mb (4.1%) were hit more than once, where the percentages in brackets are the proportions of non-coding DNA that these totals represent.

A thorough classification of all the repeat families corresponding to multi-hit regions was not attempted. A preliminary rapid clustering identified 51 putative repeat families with over 10 copies in the genome (mean copy number 20 and mean sequence length 260). Among the clusters were the Tc3, Tc7 and Tc11 elements described in Chapter 7. These families account for 20% of the multi-hit bases suggesting (by extrapolation) that there may be as many as 200 additional families. Separating these by clustering is non-trivial: statistical analyses of repeat sequences show that certain families are often found together in the genome (Section 7.4.5 of Chapter 7) which can lead to conflation of clusters.

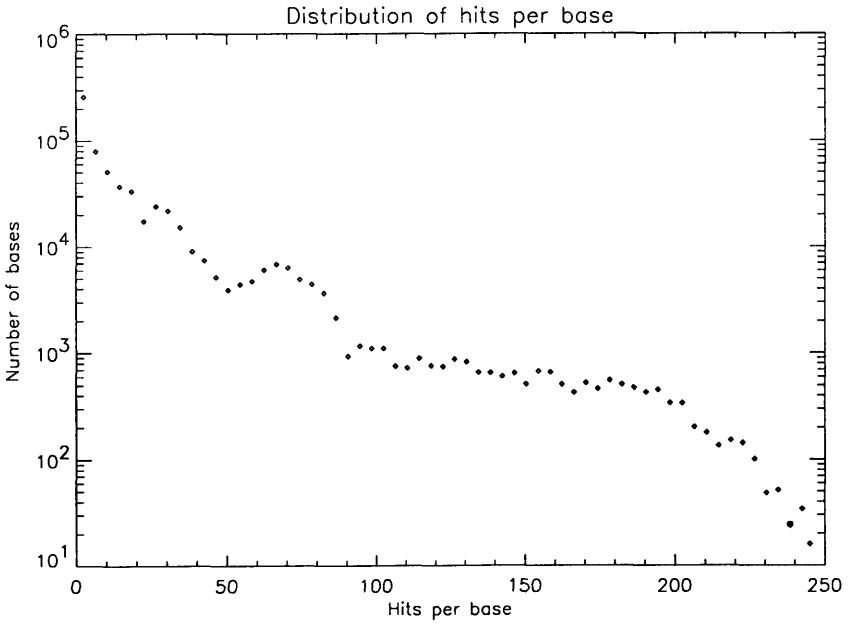

Figure 5.2: Frequency distribution of the number of times a base is involved in a high-scoring duplication.

A list of the putative new repeat families and the co-ordinates of their members is available from the Wormdup website. In total, 4520 regions that were multiply covered by the duplications data set were identified, leaving 1211 unique non-overlapping duplications in the data set. These were used for all subsequent analysis.

5.2.6 Gene duplications

In order to compare the fixation rates of non-coding and coding DNA duplications, a data set of gene duplications was independently constructed as follows. An initial search for homologies was performed for homologies between the 8065 proteins in the November 1996 release of Wormpep (the *C.elegans* database of predicted genes [SD97]) using the `blastp` program [AGM+90] with the default BLOSUM62 substitution matrix [HH92]. Gene clusters were identified on the

basis of homologies scoring over 1000 (i.e. 500 bits) or sharing over 80% sequence identity (with a score cutoff determined by the BLAST expected-hit-count parameter E=10). Minimal spanning trees for these clusters were constructed by neighbour-joining.

The above search yielded 369 multi-gene families, comprising 1035 genes. Construction of the minimal spanning tree resulted in a total of 666 duplicate pairs. This clustering is rather tight and splits up some large families; the main objective was to find a set of representative gene pairs for comparison with the non-coding pairs in Wormdup.

5.3 Statistics of duplications in Wormdup

Of the 1211 unique high-scoring duplications in Wormdup, 532 are on the same chromosome, with an approximately even split between same-orientation and inverted-orientation duplications. This suggests that many duplications are local and indeed, 52% of all same-chromosome duplicated blocks are separated by no more than 50kb. Wormdup duplications are more frequent near the ends of chromosomes, but this seems to be due to the bias induced by throwing out duplications involving genes (since genes are more densely clustered near the centre of chromosomes in *C.elegans* [ZR95]). The mean size of high-scoring duplicated blocks is 360 bases.

The number of pseudogene-like duplications in *C.elegans* (homologies between a predicted gene and a piece of non-coding DNA) was also estimated and found to be comparable to the number of non-coding duplications. The mean size of pseudogene-like duplications was also comparable to that of non-coding duplications.

5.3.1 Age distribution of duplications: the duplication fixation rate

A sequence alignment may be "dated" by finding a maximum-likelihood parameterisation of a time-dependent sequence divergence model. This is a standard technique in phylogenetic analysis and many such models have been developed; several are described in Chapter 2. The model used here was the 6-parameter model first described by Hasegawa *et al* [HKY85]; it assumes evolutionary neutrality and substitution rate constancy - the "molecular clock hypothesis".

Only the ungapped regions of the aligned duplicated blocks were used for fitting the time-dependent model. Although gap-aware models of DNA evolution exist [TKF92], their aptness is questionable. Some of these models are investigated in Chapter 6.

The Hasegawa model takes as parameters the background nucleotide composition (36% GC-content, in the case of the worm), the transition rate, the transversion rate, and the divergence time. This leaves a choice of scale for the divergence time; to fix this scale, the transition rate was set arbitrarily to 1, yielding the transversion/transition ratio k as a new parameter. This choice of scale fixes a unit of time at approximately 200 million years, though this is a general rate obtained by averaging synonymous substitution rates for a variety of phyla [LG91]. It is hard to obtain a nematode-specific figure because of the paucity of the nematode fossil record. However it is believed that the effective mutation rate has been abnormally high along the *C.elegans* lineage [Bla98].

The maximum-likelihood value of the transversion/transition ratio k and the divergence times $\{t_i\}$ of the Wormdup duplications were estimated by first choosing an empirical seed value of $k^{(0)} = 0.46$, then performing the following iteration, starting with $n = 0$: (i) fixing $k = k^{(n)}$, find the maximum likelihood times $\{t_i^{(n+1)}\}$; (ii) fixing $\{t_i\} = \{t_i^{(n+1)}\}$, find the maximum likelihood $k^{(n+1)}$. The optimisations at steps (i) and (ii) of this algorithm could be performed quickly by binary chop, since the posterior distributions of k and t_i are unimodal

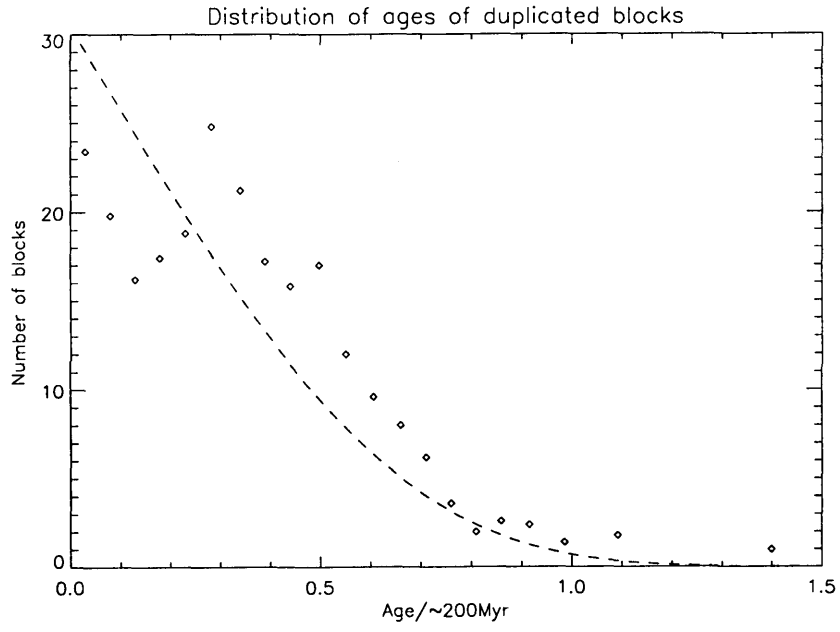

Figure 5.3: Age distribution of high-scoring duplications. The dotted line shows the hypothetical distribution that might be observed if duplication lengths were exponentially distributed, no duplications were deleted, and the only factor modulating the observed age distribution was the probability that, due to the random accumulation of substitutions and indels at the measured rates, the duplication would not score high enough to be picked up by the dynamic programming.

gamma distributions if the alignment is fixed. The algorithm was found to converge on $k = .49$ after 4 iterations.

The distribution of ages of Wormdup duplications is shown in Figure 5.3. The distribution is modulated by the probability that a duplication that old will score high enough to be picked up by the dynamic programming search (see Chapter 2). The dotted curve on Figure 5.3 illustrates this modulation assuming an initial geometric distribution of duplication lengths, which appears to be a reasonable approximation to the size distribution of recent duplications (see Figure 5.6). The observed data do not deviate plausibly from this distribution,

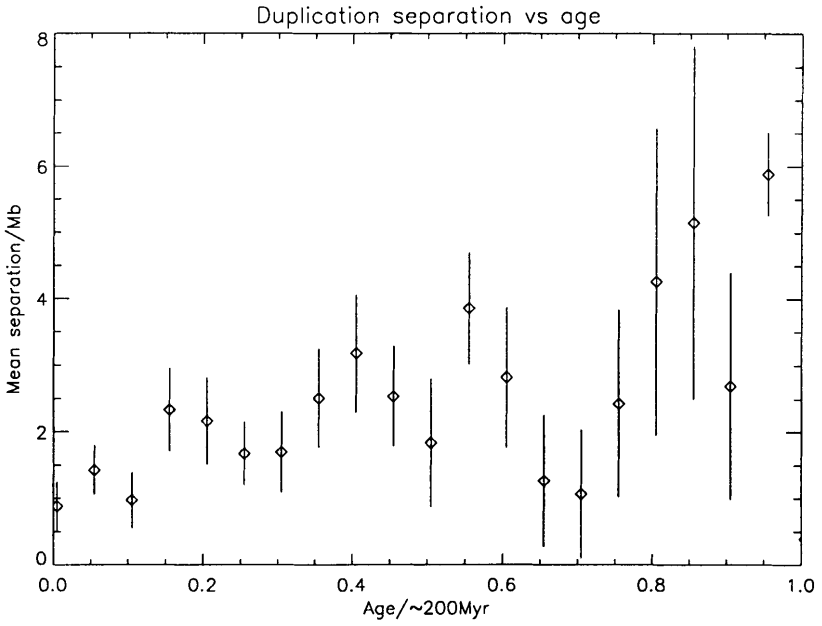

Figure 5.4: Mean separation of same-chromosome duplications plotted by age. The vertical bars indicate the standard deviation of the mean separation for each age bin.

suggesting that the rate of new duplications has remained roughly constant for the last $\sim$ 300Myr. An approximate fixation rate for high-scoring duplications can be estimated: 224 high-scoring duplications are detectable from the past 20 million years ($t \leq 0.1$), a rate of approximately 11 duplications per million years.

Figure 5.4 shows the separation of same-chromosome duplicated blocks plotted against the age of the duplications. The plot shows a slight upward trend. There is some statistical support for this; the log-odds ratio $\log \frac{\Pr[D|\mathcal{M}_1]}{\Pr[D|\mathcal{M}_0]}$ was calculated, where D is the observed data, $\mathcal{M}_0$ is a uniform Gaussian noise model for the separation data with an exponentially distributed mean (decay width 2Mb) that was integrated out and uniformly distributed standard deviation (up to 5Mb) that was optimised, and $\mathcal{M}_1$ is a linear regression model with the same

Gaussian noise and an additional exponentially-distributed gradient parameter (decay width 2Mb/t, where t is time in $\sim$ 200Myr units) that was integrated out. The log-odds score was 5.3 bits (though the hyperparameters were chosen after inspection of Figure 5.4).

This means that there is weak evidence that older duplicate blocks tend to be further apart than younger ones. Two possible explanations for this trend are offered here. The first possibility is that local duplications are being removed. One mechanism for this might be unequal crossing-over during recombination. Another might be if insertions tended to be smaller and more frequent than deletions. Although on average, the product of size and rate has to be equal for insertions and deletions if genome size is to be maintained, it is possible that relatively small, frequent insertions are balanced by relatively large, infrequent deletions (or vice versa). Large deletions in the region between a pair of duplicated blocks will be likely to delete one of the two blocks unless they are distantly separated, so the observed effect will be an excess of insertions. The second proposed explanation for the trend of older blocks to be further apart is the effect of large-scale conservative re-arrangements of the genome, such as reciprocal chromosomal translocations. Both explanations are consistent with the upward trend of Figure 5.4.

5.3.2 Length distribution of duplications: indel rates

Figure 5.5 shows the variation of average duplication size with age. The shape of this curve is mainly determined by the score cutoff. The initial downward slope is due to the accumulation of indel events with time, which modulate the length distribution (older duplications are likely to be split into smaller fragments). If the underlying distribution of gap lengths was exponential with a mean of 6 residues (corresponding to the affine gap scoring scheme used by the dynamic programming), then this effect would not be seen. This observation therefore implies that the probability of getting very large gaps is bigger than allowed for

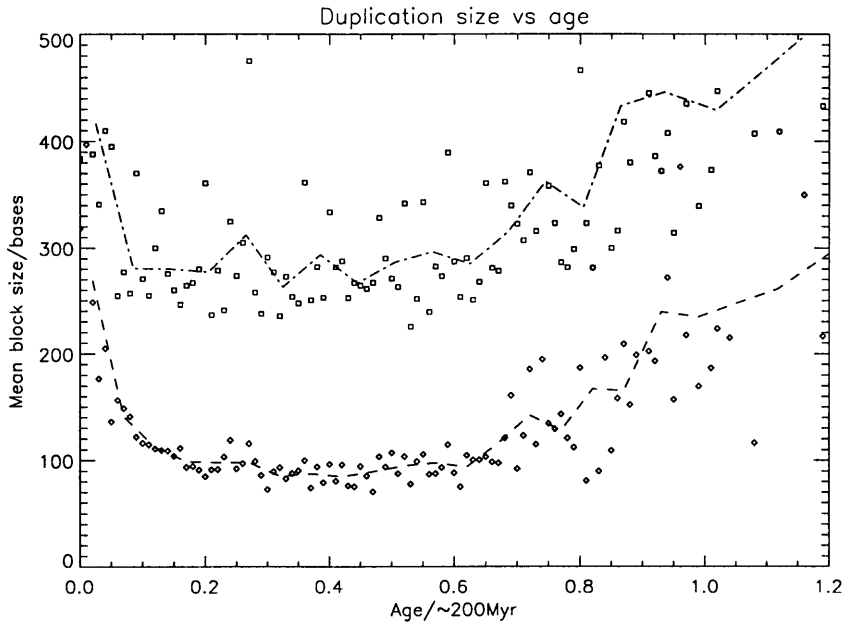

Figure 5.5: Variation of observed duplication size with age. The datapoints on the upper curve are the mean lengths of the entries in Wormdup for each age bracket. The points on the lower curve are the lengths of the constituent ungapped BLAST hits.

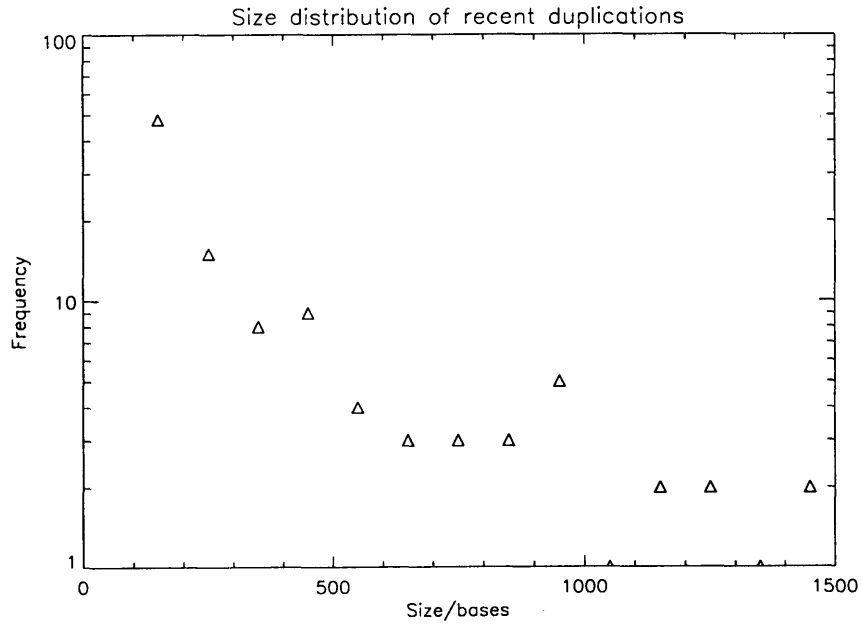

Figure 5.6: Size distribution of recent (< 10Myr) duplications. The frequency is plotted on logarithmic axes. A geometric approximation seems like a reasonable fit, although there is a hint of a broader tail suggesting that a power law distribution might be more appropriate.

by the exponential distribution. The real gap length distribution has a longer tail. Measurements of indel sizes in pseudogenes [GL95b] suggest that the gap lengths may be better modelled by e.g. a power-law distribution.

The upward turn of the graph at $t > .5$ happens because when the sequences are highly diverged, only the larger duplicated blocks stand any chance of scoring higher than the threshold for detecting hits.

A correction for the cutoff-induced bias to the observed length distribution may be derived from Bayes' theorem, and is included here for completeness although the method is not actually used. Write $\Pr[s|O,t] = \frac{\Pr[s|t]\Pr[O|s,t]}{\Pr[O|t]}$ where O is the event that a duplication is observed, s is the size of the duplication and t is the age of the duplication, which is conditioned upon throughout. An exponential approximation for the size distribution is $\Pr[s|t]\alpha \exp[-s(gt + 1/\mu)]$ where g is an indel rate and μ is the mean initial duplication size. (The actual distribution of sizes of recent duplications (younger than ~10Myr) is shown in Figure 5.6. There is a hint of a broad tail, suggesting that a power-law distribution might be slightly more appropriate than an exponential distribution, but an exponential seems like a reasonable approximation to the distribution in Figure 5.6. The distribution will tend towards an exponential with time anyway, due to fragmentation by randomly scattered indels.) The probability $\Pr[O|s,t]$ that a match of length s scores high enough to be seen may be found by approximating the match score distribution with a Gaussian (see Chapter 2), whereupon $\Pr[O|t]$ may be found numerically.

A simpler approach is to first pick a maximum age (in this case $t = 0.25$) and throw away anything older than this, then find a mean size $\hat{s}_t$ for each age t, then fit a straight line to a plot of $1/\hat{s}_t$ vs t. The gradient of this line is the rate g of fragmentary indels (i.e. indels so big they wreck the chances of putting the pieces back together again) and the y-intersect is $1/\mu$. Applying this to the Wormdup duplications gives values of $g \sim .005$ (one big indel per 40kb per million years) and $\mu \sim 400$ (the average original size of the high-scoring

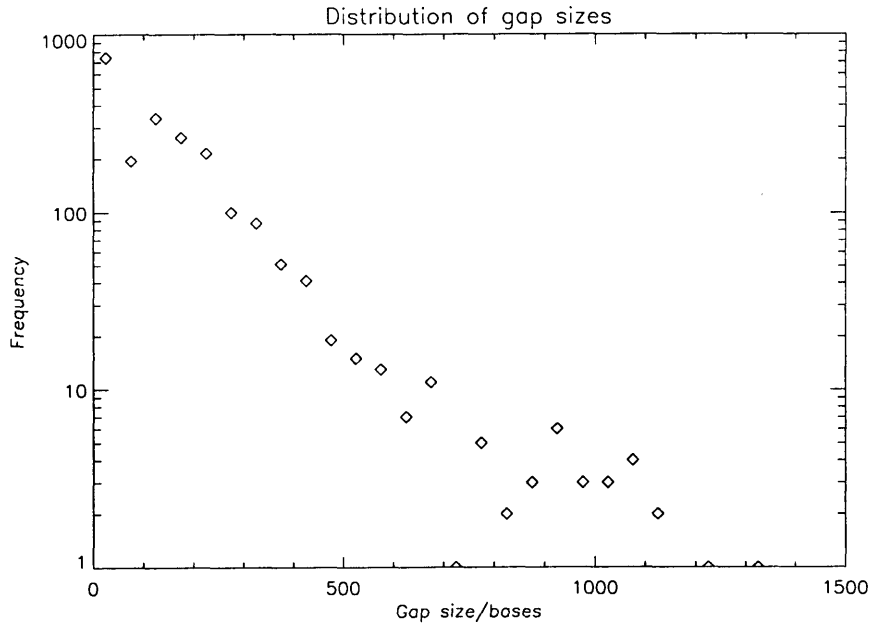

Figure 5.7: Distribution of indel sizes.

duplications was 400 bases).

A rate for small indels can also be calculated by looking at the age-length distribution of the BLAST hits from which the Wormdup duplications were derived. This gives a rate of $g \sim .03$ (one small indel per 7kb per million years) and $\mu \sim 224$.

Exactly what is meant by "big" and "small" indels? The size of indels in Wormdup appears to be exponentially distributed, with mean $\sim$ 150 bases (see Figure 5.7). This suggests that "small" indels are around 150 bases. The relative rates for small and big indels (.03 and .005) suggest that approximately 14% of indels are "big". The total rate for both big and small indels is $g_{\text{tot}} \sim .035$, or one indel per 6kb per million years.

5.4 Repetitive element-mediated duplications

Repetitive elements are known to cause duplications in a number of ways, including imperfect double-strand break repair following transposon excision [MKW91], accidental transpositon of non-transposon sequence [GL95b] and unequal crossing-over during meiotic recombination as a consequence of misalignment of adjacent copies of an element (as in Figure 1.3 of Chapter 1) [FBT+91]. In order to assess the relative importance of the latter two of these processes, a search of the *C.elegans* DNA was performed for patterns of repeat-flanked duplications whose proximity and relative orientation was indicative of repeat-mediated duplication.

The particular tool used was the `gffdp.pl` program, which implements dynamic programming using a generalised hidden Markov model of programmable structure with a pushdown stack. The `gffdp.pl` program is described (along with one of the models used for this search) in Appendix A.

In total 36 potential repeat-mediated duplications were found in the search, 22 spanning coding regions (though none of the duplicated blocks themselves intersected with coding regions). The maximum permitted separation between match segments for this search was 10kb; there are 226 matches in Wormdup that are this close together, so the hit rate to these matches was 16% (1 in 6). Nine of the 36 hits were of the form *repeat* $\rightarrow$ *match* $\rightarrow$ *repeat* $\rightarrow$ *match*, the pattern expected for unequal crossing-over events of the type shown in Figure 1.3. All 36 matches are available from the Wormdup website.

5.5 Comparison of non-coding duplications and coding duplications

Table 5.2 lists the 30 largest of the 369 clusters in the gene duplications data set whose construction is described in Section 5.2.6. Of the 666 gene pairs in the minimal spanning tree, 346 were on the same chromosome; of these, 198

were separated by under 20kb. Of the 346 same-chromosome gene duplications, 201 (58%) were oriented the same way; this proportion is even higher (64%) for those separated by under 20kb of sequence. No correlation between age and separation was found for these duplications.

The `bigdp` program (see Section A.5 of Appendix A) was used to search for blocks of genes duplicated *en masse*. Only one long-range (over 100kb) duplication involving over two pairs of genes was found, on chromosome II (pairing the three genes T05C12.3, F35C11.2 and F35C11.3 with W01C9.4, M05D6.3 and M05D6.1) and this was not very convincing, since the first gene-pair in the group is separated from the others by over 70kb. Seven long-range two-pair blocks were found.

The ages of gene duplications were estimated by fitting a time-dependent model to the observed frequencies of synonymous substitutions between the coding sequence pairs; this method is described in greater detail in Chapter 6. Most gene families are seen to have members with a wide range of ages, with notable exceptions being certain families of transposase and RNA-directed DNA polymerase proteins which probably dispersed rapidly.

If the molecular clock hypothesis holds, then the data would indicate a fixation rate of approximately 5 gene duplications per million years. 60% of these duplications involve multi-gene clusters; the rate of fixation of duplications not involving clusters is 2 per million years. Gene duplications tend to be bigger than non-coding duplications; the average *C.elegans* coding sequence is 2500 bases long, whereas the mean size of non-coding duplications in Wormdup is 400 bases. The fixation rate of gene-sized duplications in Wormdup, including pseudogenes, is estimated at 0.3 per million years. If the speed of the synonymous-substitution clock for coding DNA (the "codon clock") is the same as the speed of the substitution clock for non-coding DNA (the "background clock"), then this would indicate that gene duplications are fixed 7 times more frequently than non-coding duplications. The fraction of coding DNA in *C.elegans* is roughly

Family size	Example member	Youngest duplication	Oldest /~ 200Myr	Brief identification
32	C25A8.1	0.04	6+	Major sperm protein (msp-142)
24	F38E1.1	0	6+	transposon reverse transcriptase
20	B0280.6	0	6+	transposable element
14	ZK666.2	0	6+	DNAJ protein like
12	C50F7.5	0	6+	cuticular collagen
11	F45E1.6	0.11	6+	Histone H3
10	C03G5.5	0	1.65	unknown
9	T06C10.5	0	0.46	RNA-directed DNA polymerase
9	K03A1.6	0.05	2.17	his-10, histone-H4
8	F45F2.2	0	1.65	histone H2B
8	ZC412.2	1.96	6+	guanylate cyclase
8	F55C10.3	0	2.79	cuticular collagen
8	F55G1.10	0	1.6	histone H2A
8	T01C1.1	0	6+	reverse transcriptase
8	K07F5.9	0.11	2.36	unknown
7	F08G12.6	0	6+	transposable element Tc1 transposase
7	M03A1.5	0.08	2.17	small histidine-alanine-rich protein precursor (SHARP)
6	C08H9.7	0.39	6+	chitinase domains
6	F10F2.6	0.89	6+	C-lectin binding domain
5	T10B9.1	0.45	2.16	cytochrome P450
5	F38A5.9	0	0.06	unknown
5	F15B9.4	0.23	6+	unknown
5	ZK1248.9	0	6+	unknown
5	R04D3.1	1.27	6+	cytochrome P450
4	F52D2.3	0	0.36	transposition protein
4	T10E10.2	1.58	6+	collagen
4	C33G8.10	2.8	6+	C4-type zinc finger domain
4	F44F4.11	0	5.11	tubulin alpha-2 chain
4	ZK402.2	0.97	1.92	unknown
4	C24A3.1	0	6+	repetitive proline-rich cell wall protein

Table 5.2: The 30 largest duplicated gene families in *C.elegans*, with most recent and most ancient duplication ages also shown (dates older than 1200 million years are truncated). The clustering was tight, so that several large families were split up (e.g. cytochrome P450).

25%, so that the duplication fixation rate per base of coding DNA appears to be 20 times higher than per base of non-coding DNA.

Could the rate discrepancy between coding and non-coding duplications be due to the clocks being out of sync? That is, could the synonymous-substitution clock (or "codon clock") for coding sequences be running slower than the substitution clock ("background clock") for non-coding sequence? It is certainly easy to imagine how a wider range of mutations could affect non-coding sequence compared to coding sequence; any kind of mutation involving more than a single base will be strongly selected against in coding DNA. This would tend to make the background clock appear to run faster. On the other hand, the codon clock actually appears to run faster than the intron clock (the intron clock is based on counting the number of substitutions and indels that have accumulated inside introns; this clock is evaluated in Chapter 6). Although it is possible that the codon and intron clocks both run slower than the background clock due to selection pressures on both codons and introns, the rate for small insertions for the intron clock is similar to the background clock, suggesting an approximate correspondence. Furthermore, the observed divergence of introns is consistent with selection pressures on some introns, but not all. Variation in molecular clock rates have been reported elsewhere [GWD98] although the variations here are slightly larger.

If the fixation rate discrepancy is real, then it could be explained by positive selection pressure acting on gene duplications, greatly elevating their chances of fixation. Duplications of non-coding DNA are expected to be essentially neutral or even mildly deleterious, due to the increased DNA load. This effect would tend to elevate the relative rate of fixation of gene duplications, particularly if (as a hypothetical example) there were a selective sweep for increased dosage levels of a particular gene.

5.6 Discussion

A database of genome duplications called Wormdup has been developed from 72Mb out of the 97Mb of *C.elegans*, including a variety of tools for accessing the data set. Statistics for the database have been described, including the copy numbers of CeRep repeats and size, length and age distributions of unique duplications.

Unique non-coding duplications of the size range considered in Wormdup (mean 400 bases) become fixed at a rate of approximately 20 duplications per million years, including pseudogenes. This is a conservative estimate as multi-copy duplications were excluded. Although it is not yet clear what are the most important causes of duplication, some general trends are apparent: around half of all duplications are local in nature and no preference is shown for parallel *versus* inverted orientation. Around 1 in 6 of highly local duplications (separation <10kb) are near repetitive elements in conformations suggestive of repeat-mediated duplication, with around a quarter of these consistent with the kind of unequal crossing-over event illustrated in Figure 1.3 of Chapter 1.

Duplications do not appear to be systematically deleted on the million-year time scale, either by counterselection or by processes such as unequal crossing-over. The main process by which duplications deteriorate is stochastic accumulation of substitution and indel events. The data in Wormdup can be used to estimate the fixation rates of these kinds of small, local mutation. Fixing the rate of transitions at one substitution per 200 bases per million years, the transversion/transition ratio is estimated to be 0.49 - i.e. transitions are twice as common as transversions. Indels occur 1 every 6kb per million years; an exponential distribution with mean 150 bases models 86% of these indels. There is weak evidence that the mean separation between duplicated blocks increases with time. Two explanations have been proposed for this trend: (i) local removal of duplications, due perhaps to insertions being smaller and more frequent than deletions; and (ii) large-scale conservative re-arrangements such as reciprocal

chromosomal translocations.

The ratio between the apparent duplication fixation rates of coding and non-coding DNA is rather large at 20:1. This may mean that that the estimated non-coding rate is too conservative or that the molecular clocks are mis-calibrated. If the difference is real, it would suggest that most non-coding duplications are lost from the population. This would imply that most gene duplications that become fixed have a selective advantage.

Selection favours gene duplications that preserve orientation, even though the underlying mechanisms of duplication appear not to discriminate between preserved- or inverted-orientation duplications. A possible explanation for the preference for same-orienation duplication is that operons are used to maintain similar regulatory control over both copies of a duplicate gene pair [BS97].

The analysis shows that there are many unclassified repeat families in *C. elegans*. 48 new families were identified by a very basic clustering and it is estimated that there are around 200 more. Around 60% of these repeats are located in the outer 50% of chromosomes. A full classification and derivation of consensus sequences for repeat families would be a non-trivial project, but worthwhile if only because of the potential role of repetitive sequences in triggering duplications and the consequences for their role in evolution.

5.6.1 Availability

The Wormdup data sets are available in full online, at the following URL:

`http://www.sanger.ac.uk/Users/ihh/Wormdup/`