Genet. Res., Camb. (1999), **74**, *pp.* 23–30. *With* 4 *figures.* Printed in the United Kingdom

Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes

SVETLANA A. SHABALINA‡ AND ALEXEY S. KONDRASHOV*
Section of Ecology and Systematics, Cornell University, Ithaca, NY 14853, USA

(*Received 6 July 1998 and in revised form 9 November 1998 and 30 December 1998*)

Summary

Similarity between related genomes may carry information on selective constraint in each of them. We analysed patterns of similarity between several homologous regions of *Caenorhabditis elegans* and *C. briggsae* genomes. All homologous exons are quite similar. Alignments of introns and of intergenic sequences contain long gaps, segments where similarity is low and close to that between random sequences aligned using the same parameters, and segments of high similarity. Conservative estimates of the fractions of selectively constrained nucleotides are 72%, 17% and 18% for exons, introns and intergenic sequences, respectively. This implies that the total number of constrained nucleotides within non-coding sequences is comparable to that within coding sequences, so that at least one-third of nucleotides in *C. elegans* and *C. briggsae* genomes are under strong stabilizing selection.

1. Introduction

After an evolving lineage splits, independent mutation and random drift in the new lineages lead to divergence of their homologous DNA sequences. Different positive selection in the new lineages accelerates their divergence, while uniform negative selection due to retention of the same function in all lineages leads to selective constraint and slows the divergence down (Kimura, 1983; Ohta, 1992).

Protein-coding regions usually evolve slowly, indicating strong selective constraint (Kimura, 1983; Gillespie, 1991; Ohta, 1992), although some evolve very fast (Messier & Stewart, 1997; Swanson & Vacquier, 1998). Many instances of selective constraint were also detected in non-coding DNA (e.g. Li & Salter, 1991; Koop, 1995; Jan *et al.*, 1997), both transcribed and untranscribed, but there have been no estimates of the proportion of constrained nucleotides within the total non-coding DNA of any species. Such estimates are particularly important for multicellular eukaryotes, because non-coding DNA constitutes from approximately 70% (*Caenorhabditis elegans*) to approximately 95% (*Homo sapiens*) of their genomes (Zuckerkandl, 1992).

New data on *C. briggsae* DNA sequences (currently over 5 Mb), together with the complete sequence of *C. elegans* genome (97 Mb; *C. elegans* Sequencing Consortium, 1998) provide the first opportunity to study the pattern of similarity between the genomes of two closely related species, because the available *C. briggsae* sequences represent a more or less random sample from its genome (Marco Marra, personal communication). Here we report the analysis of the pattern of similarity within approximately 150 kb of homologous *C. briggsae* and *C. elegans* sequences.

2. Materials and methods

(i) *Sequences*

C. elegans sequences are from GenBank. *C. briggsae* sequences were obtained from the Genome Sequencing Center of Washington University (http://genome.wustl.edu/gsc/Web_pages/projects.html). Region I is stored on fosmids (cosmids) G44K07 (nucleotides 1–48276) in *C. briggsae* and on F52E4

* Corresponding author. Tel: +1 (607) 254 4221. Fax: +1 (607) 255 8088. e-mail: ask3@cornell.edu.
‡ Permanent address: Institute of Mathematical Problems in Biology, Russian Acad. Sci., Pushchino, Moscow Region, Russia, 142292.

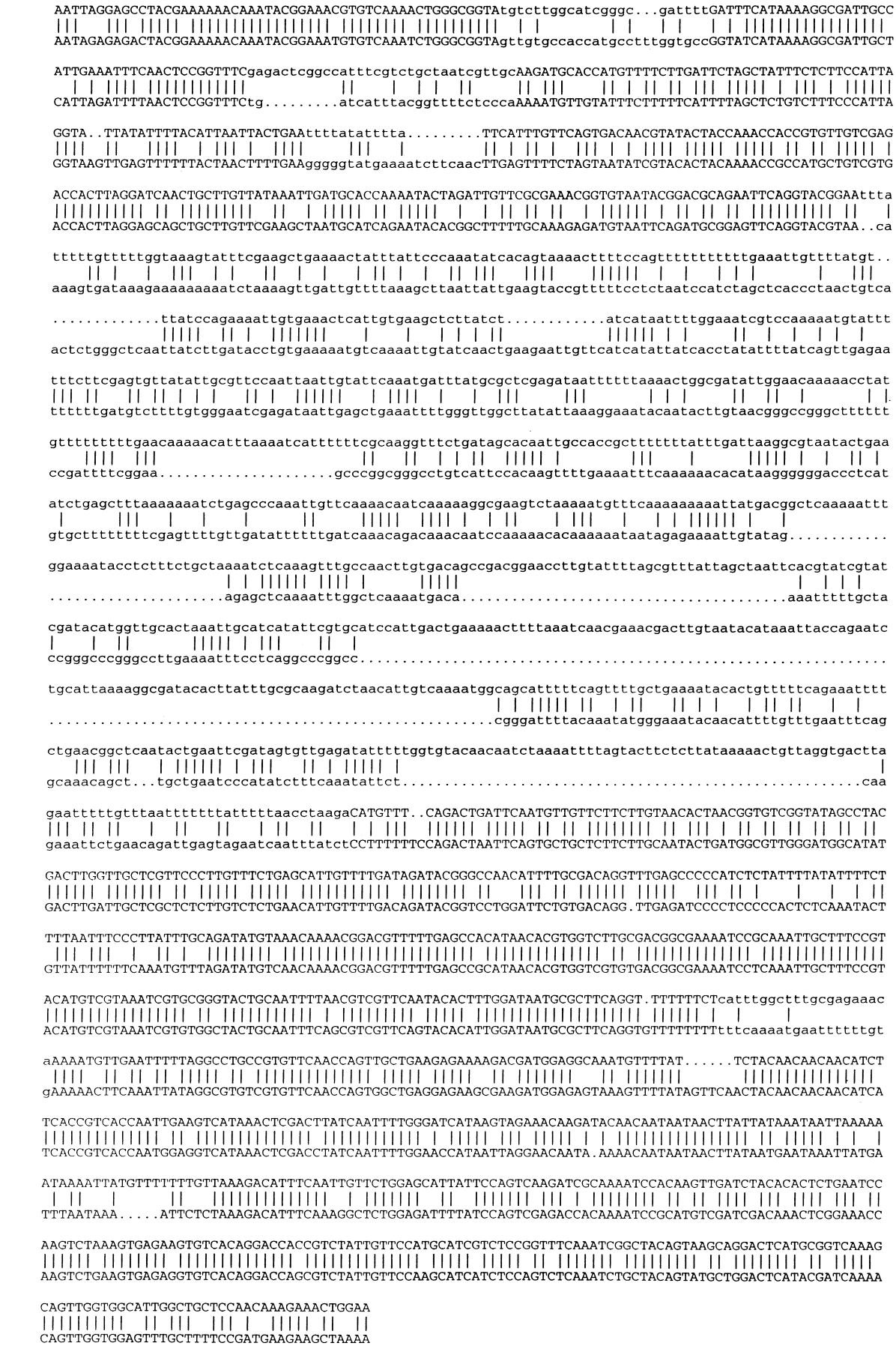

Fig. 1. For legend see facing page.

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

(9150–33796) followed by K10B3 (200–21648) in *C. elegans*. Region II is stored on G47G21 (1–40832) in *C. briggsae* and on C36B1 (1–13621) followed by F39H11 (1–18914) and K07A12 (18500–28990) in *C. elegans*. Region III is stored on G47G21 (1–43073) in *C. briggsae* and R02D5 (6788–42650) in *C. elegans*. Ten spacers are stored on CET22C1 (18015–22528) and G45023 (28386–33352), ZK381 (19610–22174) and G45C20 (17133–19420), C16C2 (7947–10001) and G41N04 (3991–5902), C18A3 (37642–42010) followed by F10C1 (1–6567) and G47F04 (20703–32358), T04H1 (10021–12600) and G46N07 (4415–7010), C41A3 (2705–4792) and G44J05 (4236–7653), T19E7 (3477–9212) and G46012 (15563–21900), C25F6 (1495–5095) and G01D9 (19240–22254), ZK632 (15657–17758) and G01C6s2 (4823–6057), and F57F5 (9555–12680) and G45L11 (9824–12275) in *C. elegans* and *C. briggsae*, respectively.

(ii) *Alignment*

All homologous sequences of *C. elegans* and *C. briggsae* analysed here contain segments of strong, unambiguous similarity interspersed with segments of much weaker similarity (see below). Due to this pattern, the following procedure was used to align these sequences. First, a pair of homologous *C. elegans* and *C. briggsae* sequences was analysed using W-Blast, which locates strongly similar segments (http://genome.wustl.edu/gsc/blast/blast_servers.html) After this, these segments were realigned individually using the program GAP (GCG), which produces an optimal alignment, i.e. an alignment having the maximal possible weight under given parameters. Segments of weak similarity, bounded by the already aligned successive segments of strong similarity, were then aligned using GAP with the same parameters. This procedure yields optimal alignment of the whole pair of sequences because interspecific homology of segments of strong similarity is unambiguous.

We have chosen those parameters that produce sensible alignments of obviously homologous sequences of *C. elegans* and *C. briggsae*. In particular, because such sequences often differ from each other by long insertions or deletions, gap length penalty must be zero or very low. The following parameters were used: match weight +1·0, mismatch weight −0·2, gap initiation weight −8·0 and gap length weight 0·0 (for both internal and end gaps).

Dr Webb Miller (personal communication) obtained, using a modified version of the algorithm described in Huang *et al.* (1990), alignments of regions I, II and III that are essentially identical to ours.

(iii) *Functional regions*

Putative locations of exons in *C. elegans* genome are indicated in Entrez data base. Alignments of the homologous sequences of the two species and program GeneScan (http://gnomic.stanford.edu/GENSCAN.html, http://CCR-081.mit.edu/GENSCAN.html) were used to find the corresponding exons in *C. briggsae* and to refine locations of exons in both species, as well as to find putative sites of initiation and termination of transcription. In most cases, alignments supported putative exons presented in Entrez. Intergenic sequences, being on average about 3000 nucleotides long, are mostly untranscribed, because the putative sites of initiation and termination of transcription could usually be found close to the first or the last exon of a gene, respectively.

(iv) *Patterns of similarity*

We subdivided each of our 13 alignments into segments reflecting the pattern of similarity in them. These 'similarity segments' may be different from those that appeared in the course of aligning. Let us define the quality of a piece of alignment as the number of matches minus the number of mismatches and the total length of gaps in it. In introns or intergenic sequences, we first found very similar 'boxes' (pieces of alignment with quality no less than 15, such that the distance between any two successive matches was below 5). After this, we attempted to expand each box, independently in both directions. The expansion proceeded until (1) the quality of the expanded box began to drop or (2) a piece of alignment with quality −10 or lower was encountered. Each expanded box constituted a segment of high similarity. A piece of alignment between two such successive segments constituted one segment of low similarity if it did not contain any gaps more than 50 nucleotides long. Otherwise, each such gap, as well as each piece of the alignment between them, was treated as a separate segment of low similarity. Each exon constituted a separate segment of high similarity. Fig. 1 shows the subdivision of a representative alignment into similarity segments. Similarity h within a segment was defined as the number of matches divided by the length of the shorter sequence within this segment. Zero similarity was attributed to gap segments.

Fig. 1. Optimal alignment of a portion of homologous intergenic regions of *C. elegans* and *C. briggsae* genomes (stored on cosmids C25F6 and G01D9, respectively) subdivided into similarity segments. Segments of high similarity are in upper-case letters and segments of low similarity are in lower-case letters.

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

(v) *Selective constraint*

The degree of similarity varies widely along *C. elegans* and *C. briggsae* genomes (e.g. Fig. 1), implying very different rates of evolution and strengths of selective constraint in different segments. Obviously, some nucleotide sites in *C. elegans* and *C. briggsae* genomes are invariant or almost invariant while others are free to evolve. Thus, we assume that there are only two classes of sites – freely evolving and invariant (Kimura, 1977; Kondrashov & Crow, 1993) – and that selective constraint within a sequence is characterized by the fraction of invariant nucleotides in it. This may be a good approximation because under a particular effective population size, the range of selection coefficients under which evolution is slowed but not completely arrested by negative selection is narrow (Gillespie, 1991; Messier & Stewart, 1997). To describe the inequality of evolution rates across sites in non-coding regions of *C. elegans* and *C. briggsae* genomes by the fraction of invariant sites clearly is more appropriate than to use, for example, a gamma distribution or a log-normal distribution, which are unimodal and contain two parameters (Waddell *et al.*, 1997).

Similarity within many segments of our alignments is the same as within alignments of random sequences (see Section 3), implying that no ancestral similarity between *C. elegans* and *C. briggsae* is left at freely evolving sites. Thus, it is impossible to estimate the number of nucleotide substitutions per freely evolving site after the divergence of *C. elegans* and *C. briggsae* (Li, 1997, chapter 4). Instead, we will estimate the fraction of invariant nucleotides. This task is complicated somewhat by the fact that such nucleotides may significantly affect the sequence alignment.

Consider a segment of alignment, with similarity h, which involves two sequences of lengths N_1 and N_2 ($N_1 \leqslant N_2$). Matches constitute the fraction $h = p_1 + c(1-p_1)$ of the shorter sequence, where p_1 is its fraction of invariant nucleotides and c is the probability that a freely evolving nucleotide in the shorter sequence corresponds to a match in the alignment. From this, $p_1 = (h-c)/(1-c)$. Because the number of invariant nucleotides must be the same in the shorter and in the longer sequences, the fraction of invariant nucleotides in the longer sequence is $p_2 = (N_1/N_2)p_1$.

In a gapless alignment of two long random sequences, c is the sum of squares of frequencies of the four nucleotides (assuming that these frequencies are the same in both sequences). This sum is 0·26 in our case (the average nucleotide frequencies, very similar in *C. elegans* and *C. briggsae*, are $\sim$0·30, $\sim$0·30, $\sim$0·20 and $\sim$0·20 for A, T, G and C, respectively). However, if gaps are introduced to increase the number of matches, c can be higher. In alignments of random sequences longer than 100 nucleotides with the above nucleotide frequencies, produced under the same parameters that were used to align actual sequences, $h = 0{\cdot}41 \pm 0{\cdot}02$ (Fig. 2). This implies $c = 0{\cdot}41$, because $p = 0$ in this case. When p_1 increases, c rapidly declines to 0·26 because invariant nucleotides begin to dictate the alignment of the freely evolving nucleotides.

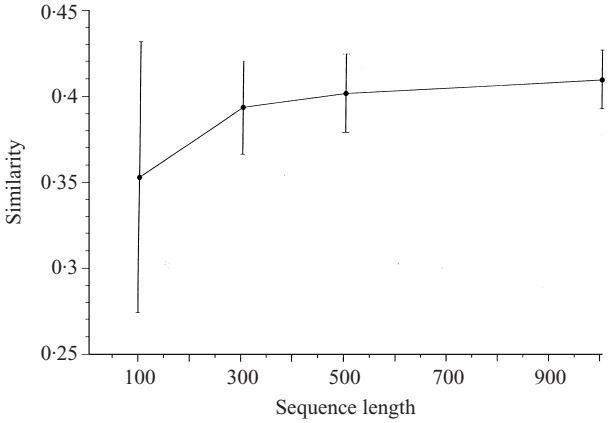

Fig. 2. Means and standard deviations of similarity h within the alignments of random sequences of equal lengths. Each point was obtained by aligning 20 independent pairs of sequences.

Aligning randomly generated sequences with the known fractions of invariant nucleotides, interspersed randomly among the others, we have found that c reaches $\sim$0·26 when $p_1 \geqslant 0{\cdot}2$. To be conservative, we attributed positive selective constraint only to segments with $h \geqslant 0{\cdot}48$, while $p_1 = p_2 = 0$ was recorded for all segments with $h < 0{\cdot}48$. Assuming $c = 0{\cdot}26$, $h \geqslant 0{\cdot}48$ implies that $p_1 \geqslant 0{\cdot}30$. Thus, we did not attempt to detect randomly interspersed invariant nucleotides if they constituted less than 30% of a segment.

3. Results

We analysed three long homologous regions of *C. elegans* and *C. briggsae* genomes, located on chromosomes X (region I), 1 (region II) and 5 (region III), as well as 10 homologous complete intergenic sequences from different parts of the genomes (two sequences on each of chromosomes X, 1, 4 and 5, and one sequence on each of chromosomes 2 and 3), each spanning the whole distance between two successive homologous genes. Polycistronic mRNAs, known in *Caenorhabditis*, usually include only closely adjacent genes (Blumenthal, 1995; Evans *et al.*, 1997). Because just one of 35 complete intergenic sequences in our sample was, in both species, shorter than 400 nucleotides, we regard these sequences as untranscribed.

Within regions I, II and III the genomes of the two species are mostly collinear, i.e. their homologous genes appear in the same order. There were, however,

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

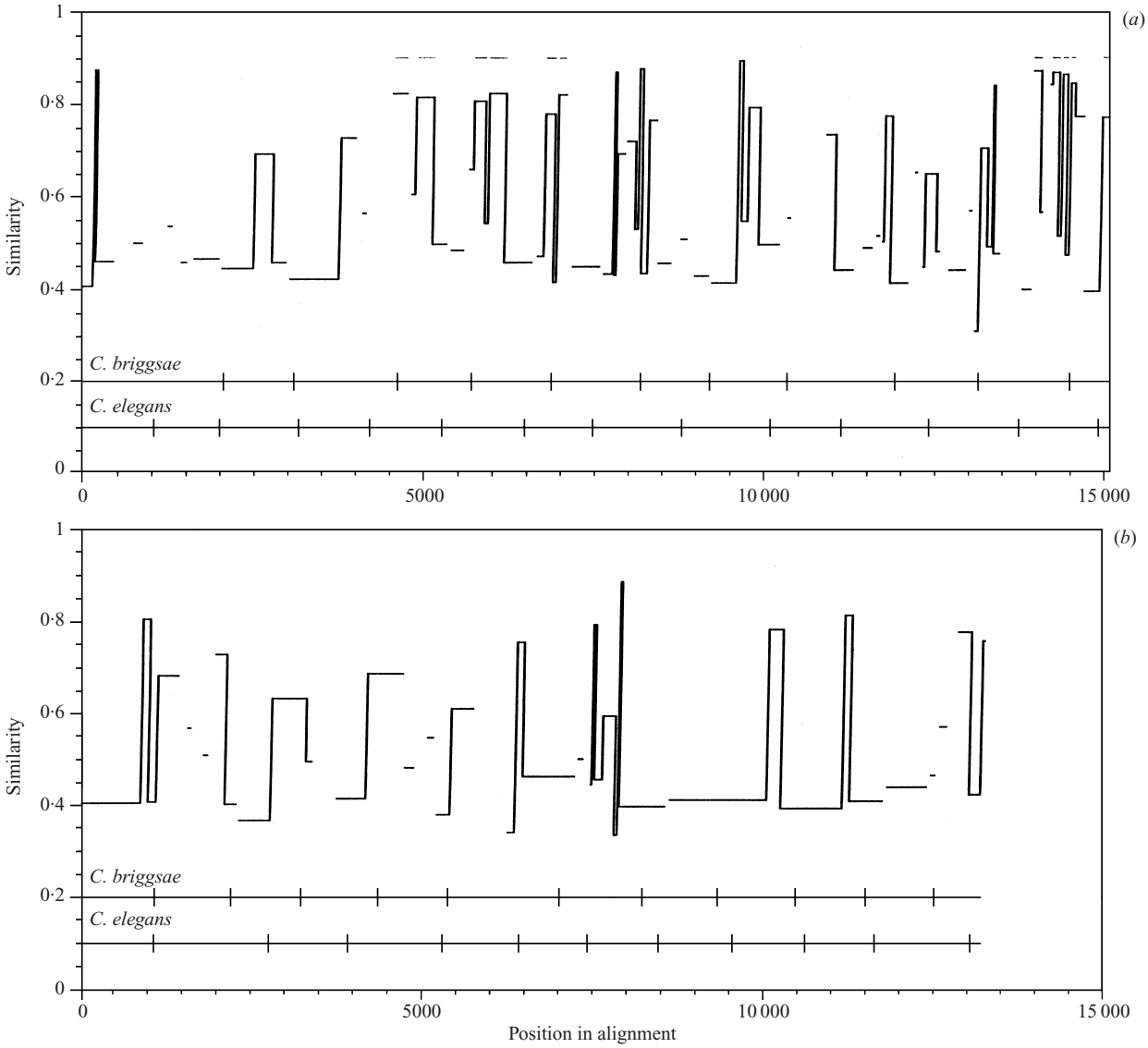

Fig. 3. Similarity in the alignments of *C. elegans* and *C. briggsae* sequences. Breaks in the level of similarity correspond to gaps more than 50 nucleotides long. Positions of every thousandth nucleotide in each aligned sequence are shown at the bottom. Due to gaps, the number of a nucleotide in a sequence lags behind its position in the alignment. (*a*) The beginning of region II, with the first two complete intergenic sequences and the first two genes (exons are shown as bars at the top). (*b*) One of 10 separate complete intergenic sequences.

two exceptions: (1) *C. elegans* has a putative extra gene between genes 7 and 8 in region I, while *C. briggsae* has a non-coding sequence between these genes, and (2) *C. elegans* has an insertion of an extra gene and a transposable element between genes 2 and 3 in region III. These regions of non-collinearity were excluded from consideration. After this, region I contained 10 putative protein-coding genes and 9 complete intergenic sequences, region II contained 11 genes and 11 intergenic sequences, and region III contained 4 genes and 4 intergenic sequences. Within each region, different genes code for rather dissimilar proteins and apparently do not constitute any clusters.

The overall lengths of the analysed exons, introns and intergenic sequences were 34917, 21500 and 101175 in *C. briggsae* and 35216, 20611 and 101447 in *C. elegans*. Thus, the average lengths of all exons of a gene, all introns of a gene and a complete intergenic sequence were 1397, 860 and 2976 in *C. briggsae* and 1409, 824 and 2984 in *C. elegans*, and exons, introns and intergenic sequences constitute approximately 28%, 16% and 56% of our sequences, respectively, which is very close to the average genomic figures for *C. elegans* (*C. elegans* Sequencing Consortium, 1998).

We aligned homologous sequences from the two genomes and analysed the patterns of similarity within these alignments. Two examples are shown in Fig. 3. Two general features of these alignments are important. First, the level of similarity fluctuates widely along the alignments: highly similar segments where

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

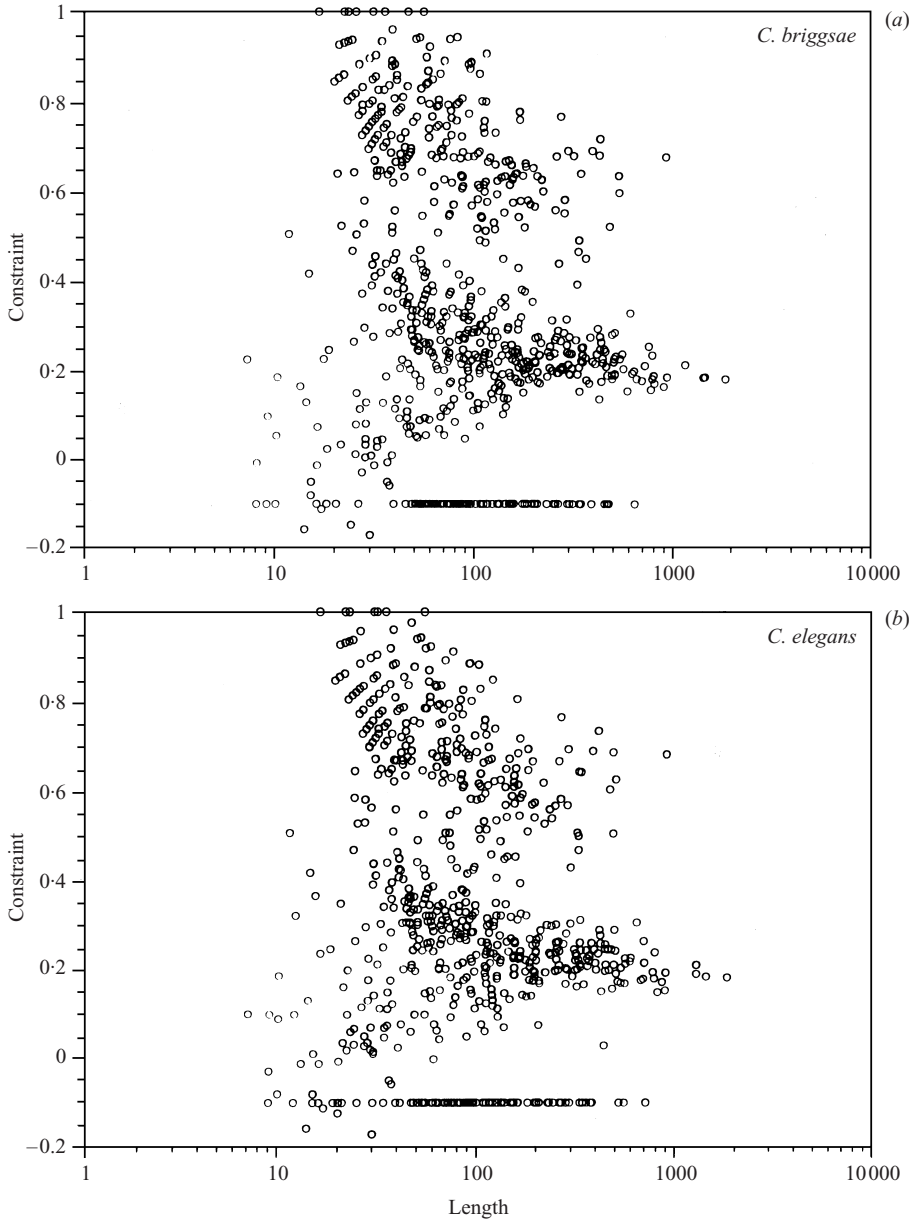

Fig. 4. Lengths and estimated selective constraints in intergenic sequences of *C. briggsae* (*a*) and *C. elegans* (*b*) corresponding to segments of different similarities defined within the alignments. Because values of constraint below 30% are unreliable, they were treated as zero constraint in our analysis. Segments aligned against gaps are shown with constraint −0·1. Other instances of negative constraint appear when similarity within a segment is below 26%.

alignment is unambiguous are separated by much less similar segments. Highly similar segments include all exons, as well as many pieces of introns and intergenic sequences. Secondly, the alignments of intergenic sequences and of some introns, but not of exons, involve substantial numbers of gaps. The total length of all alignments of intergenic sequences was 120193, i.e. 19% longer than the lengths of the aligned sequences, with 186 gaps each more than 50 nucleotides long.

Numerous segments of lower similarity within our alignments, often more than 500 nucleotides long, usually have similarity $h \sim 40\%$ (Fig. 3), i.e. the same as within alignments of random sequences with the same base composition (Fig. 2). We re-aligned 30 of such segments using several sets of parameters different from those described above and never obtained similarity substantially higher than that between random sequences aligned using the same parameters. In particular, under zero gap initiation penalty (gap length penalty was always zero; see Section 2) similarity within segments the re-aligned segments of low similarity was approximately 68%, which is not significantly different from random similarity with no gap penalties.

Optimal alignments of related genomes, obtained under appropriate parameters, can reveal similarity inherited from the common ancestor, as long as no

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

collinearity-disrupting events (inversions, duplications and transpositions) occurred in either lineage after their split. It is unlikely that such events were common in the course of independent evolution of *C. elegans* and *C. briggsae* because their gene orders are mostly the same. We conclude that, although *C. elegans* and *C. briggsae* are considered phylogenetically close (Fitch *et al.*, 1995; Sudhaus & Kiontke, 1996), there has been enough time since their last common ancestor for all traces of similarity between these species to be lost in many long regions of their genomes.

Thus, in agreement with other authors (Maduro & Pilgrim, 1996; Kuwabara, 1996), we will interpret 'over-random' similarity between many segments of *C. elegans* and *C. briggsae* genomes as the result of selective constraint. Alternatively, high similarity within a segment may reflect a locally low mutation rate. While this explanation cannot be ruled out completely at this point, it seems unlikely because synonymous site divergences for some *Caenorhabditis* genes are very high (Stenico *et al.*, 1994), although this divergence is lower for some other genes. Also, the corresponding *C. elegans* and *C. briggsae* sequences often have rather different lengths (e.g. Fig. 1) indicating that many length-changing mutations were accepted during their independent evolution. Thus, a very wide variability of mutation rates at the scale of 100–1000 nucleotides, not supported by any data, must be pervasive in *Caenorhabditis* genomes in order for the second explanation to work.

Selective constraint within a segment of a sequence, described by the fraction of invariant nucleotides it contains, was estimated from the data on interspecific similarity (see Section 2). Average selective constraint within all exons, all introns and all intergenic sequences is 0·722, 0·168 and 0·177 in *C. briggsae* and 0·715, 0·175 and 0·176 in *C. elegans*. The distributions of the mean values of selective constraint in all exons of a gene, all introns of a gene and a complete intergenic sequence has the averages and standard deviations $0{\cdot}691 \pm 0{\cdot}10$, $0{\cdot}198 \pm 0{\cdot}14$ and $0{\cdot}219 \pm 0{\cdot}12$ in *C. briggsae* and $0{\cdot}687 \pm 10$, $0{\cdot}201 \pm 0{\cdot}13$ and $0{\cdot}218 \pm 0{\cdot}14$ in *C. elegans*. The values of selective constraint within sequences that constitute all segments of alignments of intergenic sequences are presented in Fig. 4.

4. Discussion

Implications of our analysis for whole genomes of *C. elegans* and *C. briggsae* must be treated with caution, because we studied only 0·15% of them. Nevertheless, visual inspection of alignments of over 1 Mb of homologous sequences from *C. elegans* and *C. briggsae* suggests that the analysed sequences are fairly typical. If so, the fraction of functionally important nucleotides in the whole *Caenorhabditis* genome, *F*, is at least $28\% * 0{\cdot}72$ (exons) $+ 16\% * 0{\cdot}17$ (introns) $+ 56\% *$ 0·18 (intergenic sequences) = 32%, with exons containing approximately 60% of such nucleotides.

This estimate of *F* is conservative because we neglected weak constraints in introns and intergenic sequences and did not count sites that evolved due to directional selection or because some, but not all, mutations in them are neutral (e.g. if selection controls only the length, or GC-content, of a sequence). Our algorithm for defining segments, while clearly separating those that are highly conservative, may leave a substantial number of scattered constrained nucleotides within the segments from the bottom cluster (Fig. 4). The true fraction of constrained nucleotides in introns and intergenic sequences may be substantially higher than 18%.

Nevertheless, a large fraction of the *Caenorhabditis* genomes appears to be truly functionless, in particular because long gaps in the alignments suggest that many insertions and/or deletions were accepted by evolving populations. Regions of untranscribed DNA of high functional importance are interspersed among much less important or even neutral regions (Figs 1, 3).

Selective constraint in exons is primarily due to their protein-coding function, because their nucleotide sequences are less similar than the amino acid sequences derived from them (data not reported). Selective constraint in introns is at least partially due to their functioning in splicing, because it was generally higher at their edges (data not reported). Strikingly, intergenic sequences of *Caenorhabditis* consist of segments that belong to two distinct, mostly non-overlapping classes, having high versus low or absent selective constraint (Fig. 4).

The existence of such distinct classes was strongly suggested by the data on DNA–DNA hybridization involving closely related species of *Drosophila*, because while some unique sequences formed stable interspecific hybrids, others did not hybridize even under low stringency (see Powell, 1997). Of course, explicit comparison of known sequences allows one to reach more definite conclusions. The characteristic length of highly constrained segments is consistent with their involvement in DNA–protein interactions (Ptashne & Gann, 1997) and detailed small-scale analysis of the patterns of constraint within individual segments will be important for understanding functioning of non-coding DNA.

The value of *F* establishes a connection between the total diploid genomic mutation rate $T = 2\,\mu G$, where μ is the per nucleotide mutation rate and *G* is genome size, and genomic deleterious mutation rate *U*: $U = FT$ (Kondrashov & Crow, 1993). Unfortunately, there is no firm estimate of μ in *Caenorhabditis* (see Keightley & Caballero, 1997) because the reported per locus spontaneous mutation rates in *C. elegans* vary from 3×10^{-7} to $2{\cdot}5 \times 10^{-5}$ (Eide & Anderson, 1985; Schnabel *et al.*, 1991).

https://doi.org/10.1017/S0016672399003821 Published online by Cambridge University Press

So far the only multicellular organism for which a reliable estimate of μ is available is *Homo sapiens*. This estimate, $\mu \sim 2 \times 10^{-8}$ (see Kondrashov, 1998), implies $T \sim 2 \times 10^{-8} \times 7 \times 10^{9} \sim 100$ (Crow, 1997). Assuming that protein-coding sequences constitute 3–5% of the human genome and that, as in *Caenorhabditis*, comparable numbers of functionally important nucleotides occur within coding and non-coding sequences, we arrive at an estimate $F \sim 10\%$ and $U \sim 10$ for our species. Of course, this estimate of F is very tentative. We need good estimates of μ and F obtained for the same species, which is probably the best way of measuring U for the species.

We are grateful to the Genome Sequencing Center of Washington University for permission to use unpublished *C. briggsae* sequence data, to Marco Marra and Fabio Piano for helpful suggestions, to Webb Miller for independently aligning regions I, II and III, and to Alexey Spiridonov for expert programming. The work was supported by an NSF grant DEB-9417753.

References

Blumenthal, T. (1995). *Trans*-splicing and polycistronic transcription in *Caenorhabditis* elegans. *Trends in Genetics* **11**, 132–136.

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

Crow, J. F. (1997). The high spontaneous mutation rate: is it a health risk? *Proceedings of the National Academy of Sciences of the USA* **94**, 8380–8386.

Eide, D. & Anderson, P. (1985). The gene structures of spontaneous mutations affecting a *Caenorhabditis elegans* myosin heavy chain gene. *Genetics* **109**, 67–79.

Evans, D., Zorio, D., Macmorris, M., Winter, C. E., Lea, K. & Blumenthal, T. (1997). Operons and SL2 trans-splicing exist in nematodes outside the genus *Caenorhabditis*. *Proceedings of the National Academy of Sciences of the USA* **94**, 9751–9756.

Fitch, D. H. A., Bugaj-Gaweda, B. & Emmons, S. W. (1995). 18S Ribosomal RNA gene phylogeny for some Rhabditidae related to *Caenorhabditis*. *Molecular Biology and Evolution* **12**, 346–358.

Gillespie, J. H. (1991). *The Causes of Molecular Evolution*. New York: Oxford University Press.

Jan, E., Yoon, J.-W., Walterhouse, D., Iannaccone, P. & Goodwin, E. B. (1997). Conservation of the *C. elegans* tra-2 3′UTR translational control. *EMBO* (*European Molecular Biology Organization*). *Journal* **16**, 6301–6313.

Huang, X., Hardison, R. & Miller, W. (1990). A space-efficient algorithm for local similarities. *Computer Applications in the Biosciences* **6**, 373–381.

Keightley, P. D. & Caballero, A. (1997). Genomic mutation rate for lifetime reproductive output and life span in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences of the USA* **94**, 3823–3827.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kondrashov, A. S. (1998). Measuring spontaneous deleterious mutation process. *Genetica* **102/103**, 183–197.

Kondrashov, A. S. & Crow, J. F. (1993). A molecular approach to estimating the human deleterious mutation rate. *Human Mutation* **2**, 229–234.

Koop, B. F. (1995). Human and rodent DNA sequence comparisons: a mosaic model of genomic evolution. *Trends in Genetics* **11**, 367–371.

Kuwabara, P. E. I. (1996). Interspecies comparison reveals evolution of control regions in the nematode sex-determining gene *tra*-2. *Genetics* **144**, 597–607.

Li. W.-H. (1997). *Molecular Evolution*. Sunderland, Mass.: Sinauer.

Li, W.-H. & Salter, L. A. (1991). Low nucleotide diversity in man. *Genetics* **129**, 513–523.

Maduro, M. & Pilgrim, D. (1996). Conservation of function and expression of unc-119 from two *Caenorhabditis* species despite divergence of non-coding DNA. *Gene* **183**, 77–85.

Messier, W. & Stewart, C. B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154.

Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Reviews of Ecology and Systematics* **23**, 263–286.

Powell, J. R. (1997). *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. New York: Oxford University Press.

Ptashne, M. & Gann, A. (1997). Transcriptional activation by recruitment. *Nature* **386**, 569–577.

Schnabel, H., Bauer, G. & Schnabel, R. (1991). Suppressors of the organ-specific differentiation gene PHA-1 of *Caenorhabditis elegans*. *Genetics* **129**, 69–78.

Stenico, M., Lloyd, A. T. & Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Research* **22**, 2437–2446.

Sudhaus, W. & Kiontke, K. (1996). Phylogeny of *Rhabditis* subgenus *Caenorhabditis* (Rhabditidae, Nematoda). *Journal of Zoological, Systematics and Evolutionary Research* **34**, 217–233.

Swanson, W. J. & Vacquier, V. D. (1998). Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. *Science* **281**, 710–712.

Waddell, P. J., Penny, D. & Moore, T. (1997). Hadamard conjugations and modeling sequence evolution with unequal rates across sites. *Molecular Phylogenetics and Evolution* **8**, 33–50.

Zuckerkandl, E. (1992). Revisiting junk DNA. *Journal of Molecular Evolution* **34**, 259–271.