Analysis of Similarity Pattern and Selective Constraint in *C. elegans* and *C. briggsae* Genomes

Svetlana A. Shabalina shabalin@virgo.nlm.nih.gov
Alexey N. Spiridonov ahel@yahoo.com
Alexey S. Kondrashov kondrash@ray.nlm.nih.gov

National Center for Biotechnology Information, National Library of Medicine,
Bldg. 45, Room 5AN.32B, 8600 Rockville Pike, Bethesda, MD 20894, USA

Keywords: *Caenorhabditis*, DNA alignment, similarity, selective constraint

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 [2]. Protein-coding regions usually evolve slowly, indicating strong selective constraint, although some of them evolve very fast [1, 2]. Many instances of selective constraint were also detected in non-coding DNA [3], 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 ~70% (*C. elegans*) to ~95% (*H. sapiens*) of their genomes [4]. New data on *C. briggsae* DNA sequences provided the first opportunity to study the pattern of similarity and selective constraint between the genomes of two closely related species.

2 Method

We analyzed patterns of similarity between several homologous parts (~150Kb) of *C. briggsae* and *C. elegans* genomes using the following approach. First, segments of strong similarity were found using W-Blast. These segments were realigned individually using program GAP (GCG) that produces, for a pair of sequences, an optimal alignment. We have chosen parameters that produce sensible alignments of obviously homologous sequences. In particular, because such sequences often differ from each other by long insertions or deletions, gap length penalty must be 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 (both for internal and end gaps). Segments of weak similarity (usually 100–1000 nucleotides long), bounded by already aligned successive segments of strong similarity, were aligned with GAP using the same parameters. Because interspecific homology within segments of strong similarity is unambiguous, this procedure yields optimal global alignment. Alignments of the homologous sequences of the two species and program GeneScan 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.

Let us define the quality of alignment as the number of matches minus the number of mismatches and the total length of gaps in it. 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. In introns or intragenic sequences, we first found very similar "boxes" 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 alignment 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 similarty 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.

Consider a segment of alignment, with similarity h, which involves two sequences of lengths N_1 and N_2 ($N_1 \leq N_2$). Matches constitute 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, $p_1 = \frac{h-c}{1-c}$. 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 $p_2 = \frac{N_1}{N_2} p_1$.

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 same in both sequences). This sum is 0.26 for the two species, where the average frequencies are ∼0.30 for A and T and ∼0.20 for G and C. 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 under the same parameters, $h = 0.41 \pm 0.02$. This implies $c = 0.41$, 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. Aligning randomly generated sequences with the known fraction of invariant nucleotides, interspersed randomly among the others, we have found that c reaches ∼0.26 when $p_1 \geq 0.2$. To be conservative, we attribute positive selective constraint only to segments with $h \geq 0.48$, while $p_1 = p_2 = 0$ was recorded for all segments with $h < 0.48$. Assuming $c = 0.26$, $h \geq 0.48$ implies that $p_1 \geq 0.30$.

3 Results

Within the regions analyzed the genomes of the two species are mostly collinear. Exons, introns, and intergenic sequences make ∼28%, ∼16%, and ∼56% of the whole genome of each species, which is in line with previous estimates (*C. elegans* Sequencing Consortium, 1998). The degree of similarity varies widely along *C. elegans* and *C. briggsae* genomes, implying very different rates of evolution and strength of selective constraint in different segments. High similarity segments include all exons, as well as many pieces of introns and intergenic sequences. The alignment of intergenic sequences and some introns, but not of exons, involve substantial number of gaps, segments where similarity is low and close to that between random sequences aligned using the same parameters, and segments of high similarity. 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.

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*. Thus, the fraction of functionally important nucleotides in the *Caenorhabditis* genome may be at least 28%×0.72 (exons) + 16%×0.17 (introns) + 56%×0.18 (intergenic sequences) = 32%, with exons containing approximately 60% of such nucleotides. 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.

References

[1] Gillespie, J.H., *The Causes of Molecular Evolution*, Oxford University Press, 1991.

[2] Kimura, M., *The Neutral Theory of Molecular Evolution*, Cambridge University Press, 1983.

[3] Li, W.-H. and Salter, L.A., Low nucleotide diversity in man, *Genetics*, 129: 513–523, 1991.

[4] Zuckerkandl, E., Revisiting junk DNA, *J. Mol. Evol.*, 34(3):259–271, 1992.