J Mol Evol (2003) 56:371–372
DOI: 10.1007/s00239-002-2407-0

JOURNAL OF MOLECULAR EVOLUTION

Response

GC-Biased Mutation Pressure and ORF Lengthening

Antonio Marín,[1] José L. Oliver[2]

[1] Departamento de Genética, Facultad de Biología, Universidad de Sevilla, Aparta 1095, E-41080-Sevilla, Spain
[2] Departamento de Genética, Instituto de Biotecnología, Facultad de Ciencias, Universidad de Granada, E-18071-Granada, Spain

Received: 25 July 2002 / Accepted: 11 October 2002

Xia et al. (2003) discuss whether the lengths of exons in eukaryotes and of genes in prokaryotes vary and whether they do so in relation to base composition (G + C content). In the paper which generated this debate, Oliver and Marín (1996) suggested that, given the compositional AT bias of standard stop codons (TAA, TAG, and TGA), a differential density of these termination signals is expected in random DNA sequences of different base composition, and therefore the expected length of reading frames (sequence segments of sense codons flanked by in-phase stop codons) is a function of GC content. In other words, in GC-poor random sequences, the stop-codon density is expected to be higher than in GC-rich ones, and therefore the higher the GC content, the longer the expected reading frames. Empirical support for the model was sought by analyzing a sample of prokaryotic genes and a sample of eukaryotic exon data (Oliver and Marín 1996). With the model, the expected distribution of open reading frame (ORF) lengths in any random sequence with a given base composition can be computed; by comparing true ORF lengths to such random expectations, evolutionary forces involved in ORF lengthening can then be identified. Such comparisons can also be used for accurately predicting the coding content in anonymous sequences (Carpena et al. 2002).

Xia et al. reevaluate Oliver and Marín's work examining a considerably wider sample of eukaryotic exons and of genes of 68 completely sequenced prokaryotic genomes. These authors question the suggested association between base composition and ORF length mediated by differential stop-codon probability. However, Xia et al. find that, with the exception of *Mycoplasma genitalium* and *Treponema pallidum*, a positive correlation exists between ORF length and ORF GC content. Furthermore, a between-species comparison showed that the average ORF length ("genomic CDS length") and the average ORF GC content ("genomic %GC") are positively correlated among the 53 eubacterial genomes with a standard genetic code. Xia et al. acknowledge that the prediction by Oliver and Marín is largely fulfilled in prokaryotic genomes.

In this regard, we would draw attention to the "natural experiment" provided by the four Mycoplasmataceae species (i.e., *Mycoplasma genitalium*, *M. pneumoniae*, *M. pulmonis*, and *Ureaplasma urealyticum*). Such bacteria use a genetic code with only two stop codons (TAA and TAG), and interestingly the average ORF length in these species is longer than that of other bacteria with a similar GC content but using three stop codons (TAA, TAG, and TGA). Thus, as acknowledged by Xia et al., the lower probability of encountering a stop codon could be responsible for, or at least is associated with, the longer average ORF length in these species, and this

Correspondence to: José L. Oliver; *email:* oliver@ugr.es

Table 1. Elongated α-globin variants in humans that have 172 amino acids rather than the normal 141

Hemoglobin variant	Sense codon (142)	Amino acid
Constant Spring	CAA	Gln
Icaria	AAA	Lys
Koya Dora	TCA	Ser
Seal Rock	GAA	Glu
Paks	TAT	Tyr
Zurich-Altstetten	CAT	His

supports the hypothesis of an association between the ORF length and the probability of stop codons. Additional support for the model comes from the lower average ORF length in *Mycobacterium leprae* vs *M. tuberculosis*, in agreement with the AT pressure which seems to have operated on the *M. leprae* lineage (Bellgard and Gojobori 1999).

The major criticism that Xia et al. make of the model relies on eukaryotic coding sequences. According to the hypothesis, one would expect a positive correlation between the GC content and the coding part of single-exon CDSs, but Xia et al. report that such correlations in their sample of eukaryotic genes are mostly negative. We can offer no explanation for this finding, which certainly cannot be accommodated by the model in question. Nevertheless, we would like to mention that the average exon length in single-exon CDSs is disparately longer than the average exon length in multiexon genes (Tables 1 and 2 of Xia et al.), probably indicating the involvement of additional evolutionary forces. It might be that strong selective pressure is responsible for the existence of such long single-exon CDSs or that long single-exon CDSs originated by intron loss, but these explanations need further research.

Xia et al. propound that the model is applicable not to all exons, but only to the last coding exon. However, any exon must necessarily lie on a sequence free of in-frame stop codons, and it is intuitive that longer exons can exist in sequences with a lower stop-codon density, therefore the predicted correlation might extend to all exons. Actually, the results of Xia et al. show that such a positive correlation holds in most species (except yeast); nevertheless, they claim that this correlation does not support the hypothesized mechanism and they offer no alternative explanation.

However, the point of a stronger effect on the last exon is worth considering. We speculate that the finding of Xia et al. in their Table 1 (that last coding exons are longer than first or internal ones) might be due to a higher tolerance to polypeptide lengthening than to shortening, and that lengthening is more tolerated at the carboxy terminus. On this point, we present an example of hemoglobin variants which illustrates how GC-biased mutation pressure can extend a polypeptide. In humans, the globin genes of the α cluster are situated within a region subjected to GC-biased mutation pressure (Francino and Ochman 1999). A number of elongated α chains have been reported (see Table 1) which have 172 amino acids rather than the normal 141. These variants have been caused by a single-base mutation in the TAA terminal (142) codon of the α-2-globin gene and read-through of the untranslated mRNA (see Hardison et al. [2002], http://globin.cse.psu.edu, and OMIM 141850 and related entries).

In conclusion, we acknowledge that G+C variation is not likely to be the main force shaping exon lengths but current data do not exclude a role for it.

References

Bellgard MI, Gojobori T (1999) Inferring the direction of evolutionary changes of genomic base composition. Trends Genet 15:254–256

Carpena P, Bernaola-Galván P, Román-Roldán R, Oliver JL (2002) Simple and species-independent coding measure. Gene 300:97–104

Francino MP, Ochman H (1999) Isochores result from mutation not selection. Nature 400:30–31

Hardison RC, Chui DH, Giardine B, Riemer C, Patrinos GP, Anagnou N, Miller W, Wajcman H (2002) HbVar: A relational database of human hemoglobin variants and thalasemia mutation at the globin gene server. Hum Mutat 19:225–233

Oliver JL, Marín A (1996) A relationship between GC content and coding-sequence length. J Mol Evol 43:216–223

Xia X, Xie Z, Li W-H (2003) Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J Mol Evol 56:362–370