A Test of the Null Model for 5′ UTR Evolution Based on GC Content

Max Reuter, Jan Engelstädter,*[1] Pierre Fontanillas,† and Laurence D. Hurst‡*

*The Galton Laboratory, Department of Biology, University College London, London United Kingdom; †Department of Organismic and Evolutionary Biology, Harvard University; and ‡Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Eukaryotic mRNAs are headed by a stretch of noncoding sequence, the 5′ untranslated region (UTR). It has been proposed that the length of 5′ UTRs is selectively neutral and evolves under a process of stochastic destruction and recruitment of core promoter elements, combined with selection against the premature initiation of translation. We test this null model by investigating whether 5′ UTR length varies with genomic GC content, an implicit prediction of the model. Using simulations, we show that the null model predicts a positive relationship between GC content and UTR length for genes regulated by a TATA box. Although this prediction is borne out qualitatively in genomic data from yeast, fruit flies, and humans, we find marked quantitative discrepancies. We conclude that UTR length may be shaped to some degree by the forces considered in the null model but that the model fails to provide a complete explanation for UTR length evolution.

Large portions of eukaryotic genomes do not code for proteins. In order to understand how genomes evolve, it is therefore essential to study the forces shaping their noncoding parts. One component of noncoding DNA is untranslated regions (UTRs), stretches of sequence up- and downstream of coding DNA that are transcribed into mRNA but not further translated into proteins. The 5′ UTR, upstream of the coding sequence, is delimited by 2 elements. At the 5′ end lies the transcription initiation signal (TIS), a short sequence motif in the core promoter on which the transcription complex is assembled. At the 3′ end is the start codon, the first methionine codon (ATG) downstream of the transcription start site. The length of the UTR is determined by the position of these 2 elements relative to each other.

Lynch, Scofield, and Hong (2005; hereafter "LSH") proposed a null model for the evolution of 5′ UTR length (hereafter "the null model"). This model assumes that the UTR is selectively neutral and its length determined by a balance between several stochastic processes. First, UTRs are elongated when the functional TIS is destroyed by mutation and replaced with an identical sequence motif occurring by chance upstream of the destroyed motif. Second, UTRs are shortened when mutations generate new TISs within the UTR. Third, long UTRs are selected against whenever substitutions generate premature start codons within them because these lead to the production of aberrant proteins that are deleterious to the organism's fitness. Using simulations, LSH showed that the balance between mutational shifts of the TIS and selection via premature translation generates a distribution of UTR lengths that closely resembles that observed in several eukaryote species. Based on this simple test, the hypothesis that UTR length is essentially neutral cannot be rejected.

However, it has recently been shown that 5′ UTRs in yeast are longer for certain functional classes of genes (Miura et al. 2006). As this is not obviously compatible with the neutralist null model, we sought to test further this model. Here, we present an analysis of the relationship between UTR length and GC content. Although not considered by LSH, such a relationship is implicit in their null model because biased substitution rates will affect the probabilities of gaining and losing TISs and start codons. To illustrate this, consider a TIS consisting entirely of A and T. GC content affects substitutional gain and loss of such a TIS in 3 ways. First, the probability of a substitution destroying the functional TIS will increase with the degree to which substitutions are biased toward G and C (e.g., owing to a fixation bias via biased gene conversion; Galtier et al. 2001; Lercher et al. 2002). Second, the frequency (and hence proximity) of random sequences exactly matching the AT-only TIS motif upstream of the mutated element will decrease with the GC content of the sequence. Third, the probability of generating a new functional TIS within the existing UTR decreases with GC bias in substitution rate because substitution tends to generate sequences that are poor in A and T and thus dissimilar to the TIS. All 3 effects should tend to make 5′ UTRs longer in GC-rich domains than in AT-rich domains.

The 3 effects just described apply equally well to a GC-rich TIS in an AT-rich genomic environment. Thus, random gain and loss of the TIS generally tends to elongate UTRs whenever the GC content of the TIS deviates from that of the surrounding background sequence. A similar logic applies to the probability of premature start codons appearing in a UTR. Because the start codon (ATG) is AT biased, this probability decreases with GC content. Accordingly, selection against long UTRs due to premature start codons is more lenient in GC-rich regions of the genome. Mathematical expressions of these effects are given in the Supplementary Material online.

To test the null model, we generated quantitative predictions on the relationship between GC content and UTR length by simulation. In this analysis, we focused on the TATA box as TIS. The TATA box is used in a significant proportion of eukaryote promoters (e.g., Suzuki et al. 2001; Basehoar et al. 2004) and is highly AT biased (consensus TATAWAW; Benoist and Chambon 1981). Furthermore, the presence of a TATA box is known to be associated with the initiation of transcription (and hence delimitation of the 5′ UTR) at a well-defined site, compared with other TISs that have a more blurry pattern of transcription initiation (Carninci et al. 2006). The simulations were based on

[1] Present address: Institute for Integrative Biology, ETH Zurich, Switzerland

Key words: UTR, TATA box, mutation, selection.

E-mail: m.reuter@ucl.ac.uk.

Mol. Biol. Evol. 25(5):801–804. 2008
doi:10.1093/molbev/msn044
Advance Access publication February 21, 2008

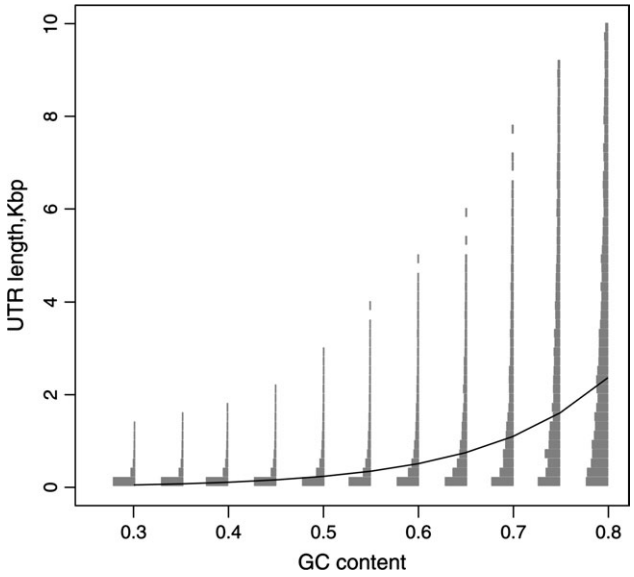

FIG. 1.—Simulated distributions of UTR length for different GC contents. The solid line depicts the predicted change in median UTR length across GC contents, obtained by fitting a linear regression through $\log_{10}$-transformed median UTR lengths ($b = 3.33$, $t_9 = 18.30$, $P < 0.0001$). This line is also shown in figure 2 and was used to compare simulations and empirical data.

a modified version of the algorithm developed by LSH, which simulates the evolution of UTR length in response to the gain and loss of TISs and translation start sites by random substitution. A detailed description of our simulation methods is given in the Supplementary Material online. Briefly, variation in GC content was taken into account by considering the effect of biased substitution rates on the turnover of transcription and translation start sites. We based our simulations on a simple model of biased substitution, in which the rate of substitution from X to Y ($r_{X \to Y}$) is proportional to the frequency of Y (f_Y) among all bases that are not X [$f_Y/(1 - f_X)$]. The base frequencies are determined by the local GC content (f_{GC}) as $f_G = f_C = f_{GC}/2$ and $f_A = f_T = (1 - f_{GC})/2$. Importantly, using this model makes our approach conservative because the substitutional bias is less biased than the local base composition ($f_Y > f_Y/(1 - f_X)$; see figure S1, Supplementary Material online), and our simulations therefore underestimate the effect of GC bias on UTR length evolution. We tested our prediction based on UTR length data from TATA-driven genes in budding yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), and humans.

As expected, the simulations predicted that UTR length increases markedly with increasing GC content (fig. 1). However, the predicted increase in UTR length with GC content was not quantitatively borne out in the empirical data (fig. 2). UTR length increased slightly with GC content in all 3 species (linear regression of log(UTR length) on GC content; yeast: slope $b = 1.44$, $t_{135} = 1.75$, $P = 0.08$; fruit flies: $b = 0.60$, $t_{419} = 2.20$, $P = 0.029$; humans: $b = 0.49$, $t_{637} = 3.73$, $P = 0.0002$); but in all cases, the slope was strikingly lower than that predicted by the null model (comparison of observed slope to predicted slope [linear regression of log-transformed

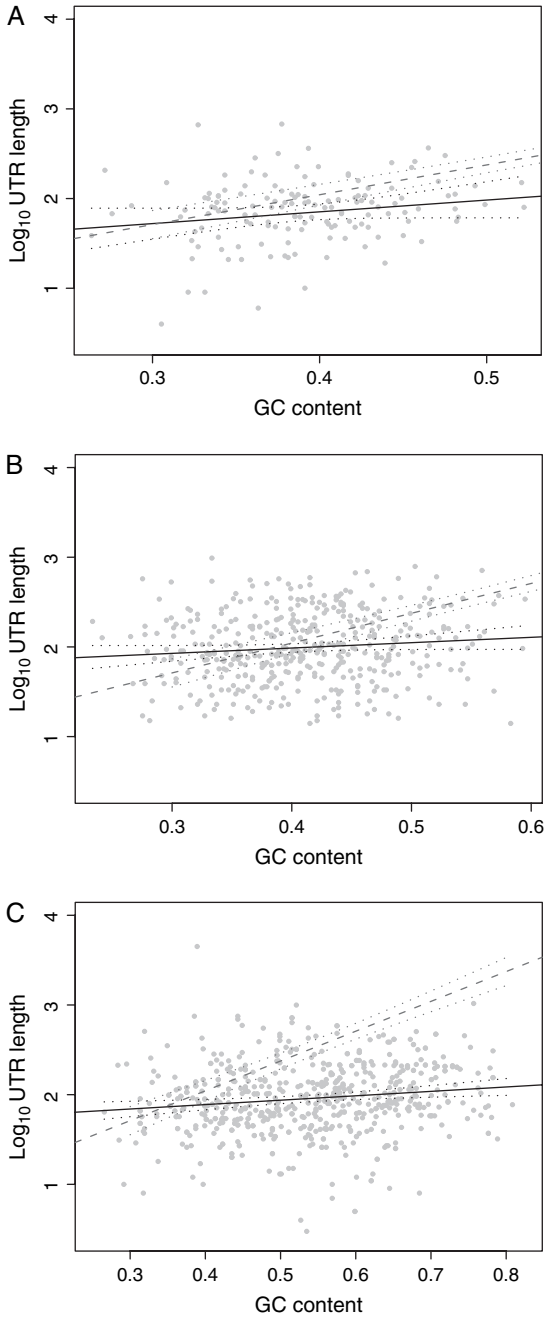

FIG. 2.—UTR length (as $\log_{10}$ of length in base pairs) as a function of GC content in TATA-driven genes of yeast (*A*), fruit flies (*B*), and humans (*C*). In each panel, the solid black line indicates a linear regression fitted through the empirical data and the dashed gray line the predicted change in UTR length under the null model (cf. fig. 1). This latter line is identical in all 3 panels; apparent differences in slope are due to the different ranges of GC content shown on the x axes. Dotted lines indicate the 95% confidence bands around regression lines.

medians of simulation data on GC content, $b = 3.33$]; yeast: $t_{135} = -3.34$, $P = 0.0005$; fruit flies: $t_{419} = -10.05$, $P < 0.0001$; humans: $t_{637} = -21.75$, $P < 0.0001$). The discrepancy revealed here between prediction and empirical data is all the more remarkable when considering the conservative way in which we implemented mutational bias in our simulations (see Supplementary Material online).

In summary, as UTR length increases with GC content in all 3 species examined and in the simulations, our results suggest that the stochastic processes invoked by the null model could play some part in the evolution of 5′ UTR length. However, the quantitative discrepancy between prediction and empirical data is too large to ignore and indicates that forces not included in the null model, acting against the random elongation of UTRs in GC-rich domains, should be considered.

What sort of forces could act against increases in UTR length? In principle, we can consider both mutational bias/neutralist and selectionist explanations. In the spirit of the former, one could hypothesize that the increase in UTR length in GC-rich regions is partly compensated by a higher deletion rate in these regions. GC-rich parts of the human genome have small introns and small intergenic distances, consistent with a deletion bias in GC-rich domains (Hardison et al. 2003; Urrutia and Hurst 2003). This effect could certainly account for some of the discrepancy between prediction and data in humans. In yeast and fruit fly, however, intergenic distance is positively correlated with the GC content of the intergenic region (yeast: Spearman's $\rho = 0.21$, $P < 0.0001$; *Drosophila*: Spearman's $\rho = 0.37$, $P < 0.0001$), suggesting that any GC-associated deletion bias does not provide a universal explanation for the fact that UTRs are not as long in GC-rich domains as expected.

As regards selectionist explanations, one potentially important factor is the cost of transcription. Highly or broadly expressed genes have been shown to have shorter noncoding sequences than less or more selectively expressed genes, a fact that has been taken as evidence for selection favoring reduced transcriptional costs (e.g., Castillo-Davis et al. 2002). Some doubts remain with respect to this interpretation, not least because of the aforementioned lower intergene distance in the vicinity of genes with small introns (Urrutia and Hurst 2003; Vinogradov 2004). Furthermore, in humans, UTR length is weakly positively correlated with expression rate in tissue-specific genes (Vinogradov 2004), arguing against a transcriptional cost model (although, note, this test does not control for GC content). Do data from yeast and fly also argue against the cost model? To examine this, we consider the relationship between 5′ UTR length and codon adaptation index (CAI), a good measure of expression rate (Coghlan and Wolfe 2000). Based on the economy hypothesis, we would expect a negative relationship between UTR length and expression rate. Data from yeast do not support this prediction because UTR length is, if anything, positively correlated with CAI (Spearman's $\rho = 0.14$, $P = 0.10$). This conclusion does not change if we repeat the analysis while simultaneously taking into account GC content, which is slightly positively correlated with CAI (Spearman's $\rho = 0.17$, $P = 0.045$). Thus, a multiple regression with GC content and ln(CAI) as independent predictors of log(UTR length) reveals that correcting for GC content removes any hint of a relationship between effect of expression rate and UTR length ($P = 0.67$), whereas the weak positive correlation between GC content and UTR is robust to control for expression rate (effect of GC content, $P = 0.04$). Comparable effects are seen in *Drosophila* (effect of ln(CAI): $P = 0.22$, effect of GC content: $P = 0.0072$).

The above evidence against the economy hypothesis is further corroborated by a comparison of UTR length between the 3 species. Based on the differences in effective population size (yeast > fruit flies > humans) and generation time (yeast < fruit flies < humans), selection due to the cost of transcription should act most strongly in yeast, less so in flies, and least in humans. This trend, however, is not borne out by the data; correcting for GC content, UTR lengths differ between species ($F_{2,1198} = 10.1$, $P < 0.0001$; ANOVA on log-transformed UTR length with covariable "GC content") and do so because fruit flies have significantly longer UTRs than the other 2 species (Tukey honestly significant differences; fly–humans: $P = 0.001$; fly–yeast: $P = 0.0008$; humans–yeast: $P = 0.3$). This and the above results suggest that the cost of transcription does not have a detectable effect on 5′ UTRs or at least not one as straightforward as favoring an ever shorter length.

Alternatively, UTR length could evolve partly in response to selective pressures arising from their function. UTRs have been shown to be implicated in cellular processes such as the regulation of transcription (see Hughes 2006 for a review). In part, this may be related to selection for nucleosome binding/nonbinding, known to be especially important for TATA-controlled genes (Ioshikhes et al. 2006). It is conceivable that stochastic changes in UTR length affect these functions in a negative way by affecting the relative position of the transcription start site to binding sites of regulatory proteins (see, e.g., Martinez-Campa et al. 2004). In addition, selection pressures on UTR length may arise from secondary mRNA structure. It has been shown that rates of translation and mRNA degradation are associated with 5′ UTR secondary structure (Ringner and Krogh 2005). Although it is not clear whether this relationship implies selection against the elongation of UTRs, it certainly suggests that random changes in 5′ UTR length are unlikely to be selectively neutral. The future will hopefully provide more insights into the role of UTRs in cellular processes that can be used to generate refined predictions about the evolution of UTR length in response to both stochastic events and different selection pressures.

Supplementary Material

An electronic appendix providing the methods used in our study is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). This document contains details on the simulations, exact mathematical expression of the probabilities for the gain and loss of TISs and start codons as a function of GC content, as well as information on how empirical data were collected.

Acknowledgments

We thank Michael Lynch for kindly providing simulation code. We thank Lars Steinmetz and Wolfgang Huber for help with yeast tiling data. This work was supported by the European Commission (Intra-European Marie Curie fellowship to M.R., Marie Curie Outgoing fellowship to P.F.), the Natural Environment Research Council, UK (Postdoctoral Fellowship NE/D009189/1 to M.R.), and the UCL Graduate School (Research Scholarship to J.E.).

Literature Cited

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. Cell. 116:699–709.

Benoist C, Chambon P. 1981. In vivo sequence requirements of the SV40 early promoter region. Nature. 290: 304–310.

Carninci P, Sandelin A, Lenhard B, et al. (41 co-authors). 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 38:626–635.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast. 16:1131–1145.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 159:907–911.

Hardison RC, Roskin KM, Yang S, et al. (18 co-authors). 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Res. 13:13–26.

Hughes TA. 2006. Regulation of gene expression by alternative untranslated regions. Trends Genet. 22:119–122.

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. 2006. Nucleosome positions predicted through comparative genomics. Nat Genet. 38:1210–1215.

Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD. 2002. The evolution of isochores: evidence from SNP frequency distributions. Genetics. 162:1805–1810.

Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. Mol Biol Evol. 22:1137–1146.

Martinez-Campa C, Politis P, Moreau JL, Kent N, Goodall J, Mellor J, Goding CR. 2004. Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. Mol Cell. 15:69–81.

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. Proc Natl Acad Sci USA. 103:17846–17851.

Ringner M, Krogh M. 2005. Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. PLoS Comput Biol. 1:585–592.

Suzuki Y, Tsunoda T, Sese J, et al. 2001. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 11:677–684.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.

Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 20:248–253.

Arndt von Haeseler, Associate Editor

Accepted February 7, 2008