A Universal Nonmonotonic Relationship between Gene Compactness and Expression Levels in Multicellular Eukaryotes

Liran Carmel*† and Eugene V. Koonin*

*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD; and †Department of Genetics, the Alexander Silberman Institute of Life Sciences, the Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem, Israel

Analysis of gene architecture and expression levels of four organisms, *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*, reveals a surprising, nonmonotonic, universal relationship between expression level and gene compactness. With increasing expression level, the genes tend at first to become longer but, from a certain level of expression, they become more and more compact, resulting in an approximate bell-shaped dependence. There are two leading hypotheses to explain the compactness of highly expressed genes. The selection hypothesis predicts that gene compactness is predominantly driven by the level of expression, whereas the genomic design hypothesis predicts that expression breadth across tissues is the driving force. We observed the connection between gene expression breadth in humans and gene compactness to be significantly weaker than the connection between expression level and compactness, a result that is compatible with the selection hypothesis but not the genome design hypothesis. The initial gene elongation with increasing expression level could be explained, at least in part, by accumulation of regulatory elements enhancing expression, in particular, in introns. This explanation is compatible with the observed positive correlation between intron density and expression level of a gene. Conversely, the trend toward increasing compactness for highly expressed genes could be caused by selection for minimization of energy and time expenditure during transcription and splicing and for increased fidelity of transcription, splicing, and/or translation that is likely to be particularly critical for highly expressed genes. Regardless of the exact nature of the forces that shape the gene architecture, we present evidence that, at least, in animals, coding and noncoding parts of genes show similar architectonic trends.

Introduction

One of the hallmarks of the architecture of eukaryotic genes is their fragmented structure (genes in pieces), where exons encoding parts of a protein are separated by noncoding introns. The fraction of intron-containing genes widely differs among eukaryotes; genes of many unicellular forms are intron poor but in multicellular eukaryotes (plants and animals), a substantial majority of the genes contain multiple introns (Rodriguez-Trelles, Tarrio, and Ayala 2006; Roy and Gilbert 2006). Generally, introns are considered to be nonfunctional, although there are many anecdotal reports on the involvement of intronic sequences in various cellular functions (Maniatis and Reed 2002; Ast 2004; Hong, Scofield, and Lynch 2006; Ying and Lin 2006; Zhao and Hamilton 2007).

It has been reported that introns in highly expressed genes in human and worm tend to be shorter than those in genes expressed at lower levels (Castillo-Davis et al. 2002). Subsequent research supported and expanded these findings by revealing the positive correlation between gene compactness and expression level in humans (Eisenberg and Levanon 2003; Urrutia and Hurst 2003; Comeron 2004; Vinogradov 2004; Chen et al. 2005; Li, Feng, and Niu 2007), mouse (Li, Feng, and Niu 2007), and microarrays (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Chen et al. 2005; Seoighe, Gehring, and Hurst 2005), and microarrays (Castillo-Davis et al. 2002; Eisenberg and Levanon 2003; Urrutia and Hurst 2003; Comeron 2004; Vinogradov 2004; Seoighe, Gehring, and Hurst 2005; Li, Feng, and Niu 2007).

Why highly expressed genes are more compact? The cell invests considerable time and energy in transcription: transcription of a single nucleotide requires at least two adenosine triphosphate (ATP) molecules (Lehninger, Nelson, and Cox 1982) and about 0.05 s (Ucker and Yamamoto 1984; Izban and Luse 1992). Thus, the compactness of highly expressed genes was attributed to selective forces that act to minimize the expenditure of energy and/or time on transcription (Castillo-Davis et al. 2002). This “selection hypothesis” received wide support from many authors (Eisenberg and Levanon 2003; Urrutia and Hurst 2003; Chen et al. 2005; Seoighe, Gehring, and Hurst 2005; Li, Feng, and Niu 2007).

However, an important observation that is not readily explained by the selection hypothesis is that intergenic regions also tend to get shorter near highly expressed genes (Urrutia and Hurst 2003; Vinogradov 2004). Although Urrutia and Hurst (2003) showed that the C1E trend remains highly significant even after controlling for this effect, Vinogradov (2004) reached the opposite conclusion. These findings led Vinogradov to propose an alternative

Key words: eukaryotic gene structure, eukaryotic gene architecture, selection on gene compactness, genomic design, intron functionality, intron density.

E-mail: carmell@cc.huji.ac.il; koonin@ncbi.nlm.nih.gov.

doi:10.1093/gbe/evp038
Advance Access publication September 22, 2009
Published by Oxford University Press 2009.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
concern, the “genomic design” hypothesis, according to which broadly expressed genes need simpler regulation, and therefore fewer regulatory elements, and less space to accommodate them than genes that are expressed only in specific conditions or tissues. The latter, narrowly expressed genes are thought to require more complex regulation and accordingly more space (both within the gene and in the flanking regions) to accommodate regulatory elements. Given that expression breadth and expression level are positively correlated, broadly expressed genes are, on average, also more highly expressed, thus highly expressed genes are more compact.

The presence of short intergenic regions near highly expressed genes correlates with the tendency of highly expressed genes to cluster along the genomic DNA (Caron et al. 2001; Lercher, Urrutia, and Hurst 2002), suggesting that local mutational bias might also contribute to the C↑E trend. This complication notwithstanding, controlling for various local characteristics like GC content, several authors demonstrated that the C↑E trend remains highly significant (Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Comeron 2004; Seoighe, Gehring, and Hurst 2005).

Surprisingly, given the consistent support of the positive correlation between gene expression and compactness in diverse organisms, the opposite trend, hereinafter denoted the C↓E trend, has been reported in flowering plants, the monocot *Oryza sativa* (rice), and the dicot *A. thaliana* (Ren et al. 2006).

These opposing observations require explanation. One obvious possibility is that the forces that shape the gene architecture in plants and animals are different (Ren et al. 2006). Else, there could be a substantial methodological difference between the work of Ren et al. (2006) and the studies that report the C↑E trend, especially considering that the study of Ren et al. is unique in using the massively parallel signature sequencing (MPSS) technology (Brenner, Johnson, et al. 2000; Brenner, Williams, et al. 2000) to measure expression levels.

Given the reported opposite trends and the uncertainty with regard to the evolutionary forces that shape the dependence between gene compactness and expression, we sought to analyze this dependence in multicellular eukaryotes within a more comprehensive framework. We examined compactness in terms of multiple-length variables, such as the length of the entire transcript, the length of the protein-coding sequence, the combined lengths of all coding and noncoding exons, and the combined lengths of the introns. In addition, we investigated a distinct characteristic of genes related to compactness, the intron density, in connection with expression. We show that, in both animals and plants, highly expressed genes are more compact than lowly expressed genes and explain how the analysis of Ren et al. (2006) might give rise to an appearance of a negative trend. The relationship between expression level and gene compactness in both plants and animals turned out to be nonmonotonic. We demonstrate, particularly for introns, a tendency to become longer (hence the genes to become less compact) with increasing expression levels, for lower expression levels. This nonmonotonic trend might result from the combined effect of opposing selective forces that make genes more compact for high levels of expression but make them less compact for lower levels of expression owing, at least, in part, to accumulation of regulatory elements in introns.

**Methods**

**Genome Annotation**

The following genome annotations were used:


Genome files were parsed using in-house Perl and Matlab scripts. Only genes for which messenger RNA (mRNA) record as well as coding sequence (CDS) record could be identified were included in the analysis, as lengths of introns, exons, and so on, were computed by comparing the mRNA structure with the CDS structure. Genes annotated as pseudogenes were removed. Genes annotated as having alternative isoforms were excluded from the analysis, unless all the isoforms differed only in their untranslated regions (UTRs). In that case, introns and exons within the coding region were analyzed, and the UTRs’ information was discarded.

**Expression Level**

Expression-level measurements were from the following sources:

- *Homo sapiens*: The data of Su et al. (2004) obtained with the Affymetrix GeneChip Human Genome U133 Array Set HG-U133A (see http://www.affymetrix.com/products/arrays/specific/hgu133.affx) were employed. In this study, the microarrays were used to measure expression in 79 different tissues, each in two repetitions. Some tissues, such as brain, blood, and bone marrow, are overrepresented in Su’s data set. Other tissues are cancerous. We have therefore selected 32 nonredundant, normal tissues for the analyses in the current study (supplementary table S5, Supplementary Material online).
- *Caenorhabditis elegans*: Combined measurements from several studies, all employing the Affymetrix GeneChip C. elegans Genome Array (see http://www.affymetrix.com/products/arrays/specific/celegans.affx) (Baugh et al. 2005; Fox et al. 2007; Falk et al. 2008), were used. Only expression in wild-type specimens was included. The final data set included 34 samples from Baugh et al. (2005), 7 samples from Fox et al. (2007), and 7 samples from Falk et al. (2008), 48 samples in total.
• *Drosophila melanogaster* : Measurements from two studies using the Affymetrix GeneChip Drosophila Genome 2.0 Array (see http://www.affymetrix.com/products/arrays/specific/fly_2.affx) (Chintapalli, Wang, and Dow 2007; Kadener et al. 2007) were analyzed. Only expression in wild-type specimens was included. The final data set included 10 samples (control only) from Kadener et al. (2007) and 44 samples from Chintapalli, Wang, and Dow (2007).

• *Arabidopsis thaliana* : The results of Meyers et al. (2004), which comprise one of the plant MPSS databases (Nakano et al. 2006), were analyzed. Expression was measured using the 17-bp signature MPSS technology. Only untreated samples were used, summing up to 10 samples in total (AP1, AP3, AGM, INS, ROS, SAP, LES, GSE, CAS, SIS [see “Library information” in http://mprss.udel.edu/at]).

Probes/target sequences that had matches in multiple genes were removed from the analysis. When multiple probes/target sequences matched a single gene, their average was taken. Only genes for which expression measurements were available were included in the analysis. The final data set consisted of 9,355 human genes, 10,071 fly genes, 15,438 nematode genes, and 14,184 Arabidopsis genes (see supplementary table S6, Supplementary Material online, for more statistics on the data set). Therefore, for a given organism, each gene in the data set has one expression-level measurement per sample. In order to put expression-level measurements from different experiments and different tissues on the same scale, we followed a technique proposed by Ren et al. (2006): For each organism, the expression data are a matrix $E_D$ with $n_g$ rows (number of genes) and $n_s$ columns (number of samples). Let $C$ be a predefined number of categories. Then, each column (sample data) is ranked into $C$ categories, such that the genes with the lowest expression levels have the value 1 and the genes with the highest expression levels have the value $C$. The number of genes in each category is kept approximately equal (as closely as possible). At the end of this process, we obtain another $n_g \times n_s$ matrix, $E_R$, with ranks replacing the original expression values. In order to obtain a single expression-level value per gene, the ranks in every row were averaged, and the result was rounded to the nearest integer. Formally, the expression of a gene $g$ is given by

\[ E(g) = \text{round} \left( \frac{1}{n_s} \sum_i E_R(g, i) \right). \]

In this work we always use $C = 30$, but the results are robust to the $C$ value (data not shown). Such definition of the expression of a gene allows combining many expression data sources but comes at the expense of using ranks. This approach allows one to analyze the general trends in the relationship between expression and compactness, as we do in this work, but masks the specific details of this relationship. We verified that adopting another strategy, based on the popular log2 expression values instead of ranks over samples, yielded qualitatively the same results (supplementary fig. S1, Supplementary Material online).

Expression Breadth

It is customary to define a threshold for making a binary decision whether a gene is present or absent in a particular tissue/condition. Such binary decisions are, by definition, somewhat arbitrary, so in this study we employed six different threshold values. These threshold values were taken as the expression levels that correspond to a certain percentile of the expression levels, ranging from 10% to 60% (see supplementary table S7, Supplementary Material online). Expression breadth was computed only for human genes and was defined as the number of tissues where the gene is called expressed under the corresponding threshold.

Segmented Regression

Segmented regression is the process of fitting data to possibly more than one linear segment (Oosterbaan 1994). We used the SegReg (http://www.waterlog.info/segreg.htm) software, which selects the most statistically significant linear model that consists of up to two linear segments. When a two-segment model is the most appropriate one, the program computes the bend point and estimates its standard error.

Results

The Relationship between Gene Compactness and Expression Level Is Universal and Nonmonotonic

We measured gene compactness using several length variables. Total lengths variables are the total transcript length, the length of the protein-coding (CDS) region, the combined lengths of all coding and noncoding exons, and the total length of introns (supplementary fig. S2, Supplementary Material online). Whenever available, we also analyzed lengths of the 3’ untranslated regions (UTRs) and the 5’UTRs. Per-gene length variables are the mean and median lengths of the exons and the introns. Expression levels were binned into 30 classes (see Methods), and the average value of each of the length variables in each of the 30 expression levels was computed across all the genes.

Qualitatively, all the organisms showed the same nonmonotonic dependence between expression level and each of the total length variables (fig. 1, supplementary fig. S3, Supplementary Material online). The compactness of genes decreases with the increasing expression level for low expression levels (the $C\uparrow E$ trend) and then increases with the increasing expression level for the high expression levels (the $C\downarrow E$ trend). To statistically validate the trend of monotonic increase of the gene length variable values followed by the monotonic decrease (hereinafter the $A$-shape), we employed segmented regression using the SegReg software (see Methods). The $A$-shape was statistically significant in all four organisms and for all the four total length variables (table 1). Notably, the $A$-shape is more significant than any monotonic trend alone. Specifically, the $A$-shape was highly significant in *A. thaliana*, and we did not obtain any support for the claim of Ren et al. (2006) that highly
expressed genes were less compact than lowly expressed genes in this plant. The lack of agreement with Ren et al. appears to stem, largely, from a series of methodological differences. Among these are the differences in selecting the tissues and genes that comprised the database and differences in the way that highly expressed genes were compared with lowly expressed ones. A detailed discussion of the possible reasons for the failure to detect the shortening of highly expressed genes by Ren et al. (2006) is available in the Supplementary Text (Supplementary Material online).

In all organisms and for all total length variables, the slope of the increasing part has a smaller absolute value than the slope of the decreasing part. In other words, the K-shape is always nonsymmetrical: The C\[E trend seen for the weakly to moderately expressed genes is relatively weak and gradual, whereas the C\[E trend that is characteristic of more highly expressed genes is considerably steeper (fig. 1).

Comparing the lengths of UTRs to expression levels yielded no significant trend (supplementary figs. S4 and S5, table S1, Supplementary Material online). It remains uncertain whether this is an indication of a genuine lack of dependence or a reflection of the poor annotation of the UTRs. Individual Introns Show A-Shape Dependence on Expression Level, whereas Individual Exons Show Monotonicity

Mean exon lengths do not display a A-shape dependence on the expression level except for a possible weak effect in Arabidopsis (fig. 2 and supplementary fig. S6 [Supplementary Material online], table 1). Human and fly show a clear monotonic decrease (a trend that is seen also in Arabidopsis except for the slight increase for lowly expressed genes), whereas, in a striking contrast, C. elegans shows a monotonic increase. As mean exon lengths reflect the combined effect of intron density and total protein lengths, the increase in average exon length with increasing expression level in C. elegans is likely to be linked to the high rate of intron loss in nematodes (Carmel, Wolf, et al. 2007).

By contrast, length–expression curves for the mean (fig. 2, table 1) and median (supplementary figs. S7 and S8, table S2, Supplementary Material online) intron length show a A-shape in all organisms. Thus, the nonmonotonic A-trend that is observed for the total length measures seems to be primarily associated with the changes in the length of introns.
Similar Dependence on Expression Level for Coding and Noncoding Regions

Both the selection and the genomic design hypotheses imply dependence between the expression level of a gene and the total length of the introns or the exons rather than the corresponding means. Our present findings reveal qualitatively the same, $K$-shape of the length–expression curve for all employed total length variables. In itself, this universal form of the dependence does not, of course, indicate that the details of this dependence, in particular, for exons and introns, are the same.

To further compare the dependences of the coding and noncoding regions on expression level, we computed for all the genes in each organism the correlation between the CDS length and the total intron length (the raw correlation) and found it to be positive and significant in all cases (table 2). If this correlation ensues from the two variables being

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The Nonmonotonic Dependence between Gene Length Variables and Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bend Point</td>
</tr>
<tr>
<td>Total transcript length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
<tr>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>CDS length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
<tr>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>Total exon length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
<tr>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>Total intron length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
<tr>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>Mean exon length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
<tr>
<td></td>
<td>AT</td>
</tr>
<tr>
<td>Mean intron length</td>
<td>HS</td>
</tr>
<tr>
<td></td>
<td>DM</td>
</tr>
<tr>
<td></td>
<td>CE</td>
</tr>
</tbody>
</table>

The results of segmented regression applied to the data in figure 1. Bend point: The expression-level category that is the border between the increasing and the decreasing parts, as decided by SegReg. The ± symbol indicates standard error (SE). Whenever the curve does not show $A$-shape, the bend point is not reported. Left fraction: The fraction of genes in the increasing part of the curve. Left slope: The slope of the left part of the curve, as computed by SegReg. The ± symbol indicates SE. Right slope: The slope of the right part of the curve, as computed by SegReg. The ± symbol indicates SE. If not otherwise indicated, SegReg found statistical support in favor of two joint linear segments.

<sup>a</sup> SegReg found statistical support for two disjoint linear segments.

<sup>b</sup> Computation was made on median values.

<sup>c</sup> Computation failed in SegReg (software crashed).

![Fig. 2.—Mean lengths of exons and introns as functions of expression-level category. All designations are as in figure 1.](image-url)
Table 2

<table>
<thead>
<tr>
<th>Raw versus Binned Correlation Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Raw Correlation</strong></td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>HS</td>
</tr>
<tr>
<td>DM</td>
</tr>
<tr>
<td>CE</td>
</tr>
<tr>
<td>AT</td>
</tr>
</tbody>
</table>

Raw correlation: The correlation was computed using all the genes. Binned correlation: The correlation was computed between the mean values across the expression-level categories. P value: The significance of the correlation change due to the binning was computed using 1,000 bootstrap repetitions.

In agreement with previous observations (Eisenberg and Levanon 2003), gene expression breadth (i.e., the number of tissues in which a gene is expressed above a threshold level) in human is positively and highly significantly correlated with expression level (supplementary table S3, Supplementary Material online). However, for all total length variables and for all threshold values, the segmented regression against breadth explained less variance than the segmented regression against expression level (supplementary fig. S9, table S3, Supplementary Material online).

Differences in GC Content Cannot Explain the Observed Relationship between Gene Compactness and Expression Level

To determine whether our findings could be affected by local mutational biases that are known to depend on the base pair composition of DNA (Duret, Mouchiroud, and Gautier 1995), we computed average GC values for all exons, all introns, and the protein-coding sequences in human. As with expression breadth, GC content did not show any significant trend, neither concave nor monotonic, with the expression level (supplementary fig. S10, table S4, Supplementary Material online).

Intron Density Tends to Increase with Expression Level

We observed previously that introns appear to be gained at higher rates in evolutionarily highly conserved genes than in faster evolving genes, with the implication that the intron density of a gene might be functionally relevant (Carmel, Rogozin, et al. 2007). Similarly, it was shown that ancient eukaryotic genes, on average, have a higher intron density than genes of apparent more recent origin (Wolf et al. 2009). We computed intron density for the four analyzed organisms as a function of expression-level category for all four organisms (fig. 3). In accord with the previous findings (Carmel, Rogozin, et al. 2007), there was a significant positive correlation between expression level and intron density in three of the four analyzed organisms (table 2). Only *C. elegans* showed the opposite trend, as previously reported by Fahey and Higgins (2007), possibly, owing to its atypically high intron loss rate (Carmel, Wolf, et al. 2007). Our results agree with those of Fahey and Higgins for the nematode but not for *D. melanogaster* for which they report the same trend as in *C. elegans*, whereas we observe a trend consistent with that seen in human and Arabidopsis (fig. 3).
The increase in intron density in parallel with increased expression is likely to reflect widespread functional importance of introns in eukaryotic genomes. This relationship is further supported by the substantial increase of the correlation coefficient when computed using the mean value of the intron density in each expression-level class (table 2).

Discussion

Comparative analysis of the connections between gene architecture and expression levels in four multicellular eukaryotes, *H. sapiens*, *C. elegans*, *D. melanogaster*, and *A. thaliana*, revealed a universal qualitative relationship between expression level and compactness (fig. 4). Unexpectedly, this trend is nonmonotonic whereby, up to a certain limit, genes become less compact with the increase of the expression level (the C↓E trend), but with further increase in expression, become more compact again (the C↑E trend). Thus, somewhat paradoxically, eukaryotic genes with the lowest expression level tend to be about as compact as the most highly expressed genes, whereas moderately expressed genes are the least compact ones.

Numerous results of comparative genomics and evolutionary systems biology indicate that gene evolution is shaped by complex interaction of many factors, some of which are general, whereas others are gene-specific, and some are selective, whereas others are neutral (Nei 2005; Lynch 2007; Ellegren 2008; Koonin 2009). The existence of a universal, nonmonotonic dependence between gene compactness and expression level seems to suggest that gene architecture depends on the interplay of several factors at least some of which are selective. The bending of all the expression–length curves at the transition from moderate to high expression levels (fig. 1) seems to be incompatible with a single underlying mechanistic cause. Furthermore, we identified a significant positive correlation between intron density and expression level, in all analyzed organisms except for *C. elegans*. Combined with the previous observations on the higher intron density and increased intron gain rate in evolutionarily conserved genes (Carmel, Rogozin, et al. 2007; Wolf et al. 2009) and the strong positive correlation between expression level and evolutionary conservation (Pal, Papp, and Hurst 2001; Koonin and Wolf 2006; Drummond and Wilke 2008; Wolf et al. 2009), this link is compatible with a functional significance of introns in expression, perhaps, leading to selection for intron accumulation. Indeed, multiple pieces of evidence in support of the involvement of introns in the regulation of expression of individual genes have been reported (Le Hir, Nott, and Moore 2003; Nott, Meislin, and Moore 2003; Moore 2005).

In contrast to the highly significant (even if complex) link between expression level and gene compactness, we found a much weaker dependence between compactness and expression breadth across animal tissues. This observation directly falsifies the principal prediction of the genome design hypothesis (Vinogradov 2004).

Thus, the results of the present analysis favor the selection hypothesis as the principal explanation why highly expressed genes are more compact than genes expressed at a lower level (the C↑E trend). What factors underpin this selection remains an open question. The original explanation implicated the minimization of ATP expenditure and time spent on the expression (transcription combined with transcript maturation) of the given gene (Castillo-Davis et al. 2002). These remain potentially relevant factors, but their ultimate importance could be questioned considering that the expenditure of both energy and time during transcription is small compared with that during translation, whereas the connection between intron length and splicing rate is uncertain. An alternative selective factor could be the fidelity of transcription, splicing, and in the case of exon sequences, translation. Evolution of both nonsynonymous and synonymous sites in protein-coding sequences, in particular, those of highly expressed genes, appears to be substantially constrained by selection for robustness to mistranslation-induced protein misfolding (Drummond and Wilke 2008; Wolf, Wolf, and Koonin 2008). Selection for short exons in highly
expressed genes could be part of the same general trend. Similarly, selection for short introns might have to do with the minimization of splicing errors that result in the formation of partially spliced transcripts which could be not only a waste of time and energy but also, perhaps, more importantly, a source of potentially toxic, misfolded proteins.

Explaining the negative trend (C↓E) between gene compactness and expression that is observed for genes that are expressed at lower levels seems to be more difficult than interpreting the positive trend. Highly expressed genes, on average, encode more evolutionarily conserved proteins than lowly expressed genes (Koonin and Wolf 2006; Drummond and Wilke 2008). This trend is a key aspect of the “status” of a gene in an organism, a proxy of its biological importance, broadly understood (Koonin and Wolf 2006). This trend is a key aspect of the “status” of a gene in an organism, a proxy of its biological importance, broadly understood (Koonin and Wolf 2006). The increased total length of introns in moderately expressed genes compared with lowly expressed genes, in part, is due to the increased intron density. However, the (C↓E) trend is clearly seen even in *C. elegans*, where intron density drops with the increase in expression level. The most relevant phenomenon to account for the (C↓E) trend might be the accumulation, in introns, of regulatory elements contributing to expression (Le Hir, Nott, and Moore 2003; Nott, Meislin, and Moore 2003; Moore 2005). Thus, perhaps paradoxically, a version of the genomic design hypothesis could be relevant to explain the negative correlation between expression and gene compactness that is seen on the low end of the expression scale.

The most surprising finding of this analysis seems to be the universality of the dependence between expression level and gene compactness in all analyzed organisms (figs. 1 and 4), in spite of the major differences in their gene architectures. As a case in point, the trend between the total gene length and expression level is essentially the same in humans and Arabidopsis despite the fact that, in humans, the total length of introns is much greater than that of exons, whereas the opposite holds for Arabidopsis. Thus, the evolutionary factors that shape the dependence between gene compactness and expression appear to be more fundamental than those that govern the evolution of gene architecture (determined, primarily, by intron lengths), in agreement with the growing evidence that expression is one of the major determinants of gene evolution (Drummond and Wilke 2008; Wolf, Wolf, and Koonin 2008). The effects of gene compactness on expression are certainly amenable to further comparative-genomic and experimental analyses, which should lead to insights into the nature of the universal relationship described here.

**Supplementary Material**

Supplementary figures S1–S10, tables S1–S7, and supplementary text are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/)

**Acknowledgments**

The authors would like to thank Miklós Csúrós and Igor Rogozin for helpful discussions. The authors’ research is supported by the intramural funds of the US Department of Health and Human Services (National Library of Medicine).

**Literature Cited**


Kadener S, Stoleru D, McDonald M, Nawathean P, Rosbash M. 2007. Clockwork Orange is a transcriptional repressor and

Kateryna Makova, Associate Editor
Accepted September 19, 2009