Abundance of dinucleotide repeats and gene expression are inversely correlated: a role for gene function in addition to intron length

Vinee K. Sharma, Navleen Kumar, Samir K. Brahmachari, and Srinivasan Ramachandran

G. N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Delhi, India

Submitted 22 August 2006; accepted in final form 30 May 2007

EUKARYOTIC TRANSCRIPTION is a slow, costly, and complex process involving the interaction of several proteins and regulatory sequences (8). About 18–25 nucleotides are transcribed per second incurring an expenditure of at least two ATP molecules per transcribed nucleotide (22–24). Given this inherently demanding process, several intrinsic factors operating at the level of template DNA have been identified that serve to facilitate the attainment of high transcriptional levels in eukaryotes. Among these, gene length appears to play a significant role. It was observed that highly and broadly transcribed genes (across a wide variety of tissues) are generally shorter compared with weakly transcribed genes (7, 11, 47, 48). These observations suggest that selection for short length in highly transcribed genes is perhaps driven by a necessity to minimize energy expenditure during transcription. Highly and broadly transcribed genes generally carry out housekeeping functions essential for maintenance of cellular physiology. Furthermore, Akashi and Gojobori (3) pointed out that ribosomal proteins, which are housekeeping, use higher frequency of less costly amino acids. Thus minimization of the energy budgets invested in gene expression occurs at all levels in a cell (47).

Although length is an important determinant of gene transcription, this by itself is insufficient. Eukaryotic DNA is hierarchically organized, and therefore other factors in addition to length must play important roles in controlling the level of gene transcription. It was observed that broadly transcribed genes tend to be clustered in the genome (25–26). Similarly, Lercher et al. (26) and Marin et al. (30) observed that highly transcribed genes have elevated (G+C) content. Since a negative correlation was observed between the nucleosome formation potential and the (G+C) content of genes (48), these observations taken together suggest that genes that are clustered and highly and broadly transcribed are located in regions of “open chromatin.”

The length, G+C content and the nucleosome formation potential of gene sequences are intrinsic properties stemming from sequence patterns and organization. Another important intrinsic property of the sequence of a gene is its spatial structure. Several years ago it was pointed out that transcription and supercoiling of the template DNA are interrelated, and therefore, the topology adopted by the template DNA is likely to affect transcription (17, 27, 49, 51). The normal structure of DNA in the cell is the B-form, a right-handed helix with 10.4 base pairs per turn. But small segments consisting of unusual compositional characteristics such as alternating purine-pyrimidine simple sequences (CG)n or (TG/CA)n display propensity to adopt a Z-form, a left-handed helix with 12 base pairs per turn under conditions close to physiological (18, 31). Such a transition from a B-helical to a Z-form occurs in the negatively supercoiled DNA due to presence of these repeats and affects the movement of RNA polymerase (Fig. 1) (34). In vitro experiments show that this type of inhibition of transcription is most pronounced at CG type sequence motifs (34).

Among all the dinucleotide repeats, (CG)n repeats are under represented in the human genome, whereas (TG/CA)n repeats are the most abundant (50%) followed by (AT)n repeats (35%) and (GA/TC)n repeats (15%) (12, 13, 39). The remaining dinucleotide repeats show very low distribution. In recent years, experimental evidence has been accumulating on the role of uninterrupted (TG/CA)n repeats as cis-modulators of transcription (2, 10, 16, 31, 32, 35, 39–43). The direction and extent of transcriptional modulation by (TG/CA)n repeats vary among genes. However, in most cases, it was observed that (TG/CA)n repeats of length n ≥ 12 units exert a downregulatory effect on transcription. Moreover, this negative modula-
tory effect increases with increasing length of repeats (Fig. 1). The influence of (TG/CA)_n repeats most likely arises from the adoption of a non-B form of the template DNA undergoing transcription. Among the other dinucleotide repeats, (TA)_n and (GA/TC)_n repeats are also shown to have influence on gene expression (5, 6, 9, 28). In this scenario, we focus on the examination of association or abundance of four basic types of dinucleotide repeats (TG/CA)_n, (AT)_n, (GA/TC)_n, and (GC)_n in human genes expressed at different levels. The remaining dinucleotide repeat types are equivalent to these four basic types (13). Such an analysis can illuminate another intrinsic feature of the template DNA, namely, the correlation of transcriptional activity with abundance of uninterrupted dinucleotide repeats.

METHODS

Microarray Datasets

Dataset A. Normalized human gene expression data obtained from the Affymetrix array (HG U95Av2) experiments using blood leukocytes from 13 human individuals including monozygotic twins is available at Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo) under following accession numbers: GSM14477, GSM14478, GSM14479, GSM14480, GSM14481, GSM14482, GSM14483, GSM14484, GSM20645, GSM29053, GSM29054, GSM29055, GSM29056, GSM29057, and GSM29058 (38). Only genes with present “P” call were considered.

Dataset B. Normalized human gene expression data from the Affymetrix array (HG U95A) experiments were obtained from the Gene Expression Atlas database (http://expression.gnf.org), which contains information on gene expression from 46 different human tissues, organs, and cell lines (44). Only genes with average difference values >200 units were considered (44).

Dataset C. Normalized human gene expression data from Affymetrix array (HG U133A) experiments were obtained from the GNF SymAtlas database (http://symatlas.gnf.org/SymAtlas/), which contains information on gene expression from 79 different human tissues, organs, and cell lines (45). Only genes with present “P” call were considered.

Filtering and Binning

A sum of 212 genes from dataset A, 192 genes from dataset B, and 188 genes from dataset C, whose expression varied in monozygotic twins (38), were filtered out. Five additional genes were also removed from dataset B because of ambiguous annotations. The remaining dataset A had 5,015 genes, dataset B had 6,650 genes, and dataset C had 11,017 genes. The signal intensities (datasets A, C) and average difference values (dataset B) were transformed by taking logarithms to the base 10. The log-transformed values were averaged across all the samples. This value, corresponding to log of geometric mean, was considered as the average transcriptional level of a given gene (47). Genes were arranged in increasing order of transcriptional levels, and 5% of the total number of genes was sliced out for each bin. These bins were numbered in an increasing order of bin average of average transcriptional levels of genes. In dataset A, the bins from 1–15 contained 251 genes each, and the remaining bins 16–20 contained 250 genes each. In dataset B, bins 1–10 contained 333 genes each, and the remaining bins 11–20 contained 332 genes. In dataset C, the bins 1–17 contained 551 genes each, and the remaining bins 18–20 contained 550 genes. The bin average value of transcription of genes in each bin was used for further analysis.

Housekeeping Genes

The datasets of human housekeeping genes were obtained from (HuGEIndex) (19) and Eisenberg and Levanon (11). Out of 451 housekeeping genes from HuGEIndex, 418 had known HGNC gene symbols. Out of 575 housekeeping genes identified by Eisenberg and Levanon, 565 had known HGNC gene symbols. We repeated the binning exercise after removing the housekeeping genes belonging to the three datasets. A total of 285, 279, and 306 HuGEIndex housekeeping genes were removed from datasets A, B, and C, respectively. Similarly a total of 407, 408, and 412 housekeeping genes of the
Eisenberg and Levanon dataset were removed from datasets A, B, and C, respectively.

Sequence Retrieval and Mapping of (TG/CA)n=12 Repeats

Sequences of human genes (latest updated as of Nov.–Dec. 2006) were retrieved from National Center for Biotechnology Information (50). The start and end of a gene are considered as the nucleotide sequence from the 5’-end of first exon to the 3’-end of last exon including untranslated regions at 5’-end and 3’-end (UTRs). If alternate splicing was reported, the gene length considered was the longest, including all alternatively spliced products for that gene (48).

Uninterrupted (TG/CA)n=12 repeats (type II and type III) were identified as described previously (39, 41). Repeats of these types have been shown to modulate transcription (2, 10, 16, 31, 32, 35, 39–43). Furthermore, repeats of this length have also been shown to have preferential binding to nuclear factors compared with short repeats (14) and can also stimulate mRNA splicing (20–21). Similarly, uninterrupted (AT)n=12 and (GA/TC)n=12 repeats were also identified to analyze the association of these repeats with gene expression (5, 6, 9, 28). All repeats were scored in the intragenic region (exons, introns, and UTRs) only.

Functional Classification of Gene Families for Comparative Analysis

The genes belonging to bins that showed significantly higher or lower than expected proportion of (TG/CA)n=12 repeats were aligned against the KOG database (46) by stand-alone version of BLASTX (50), and the best hit for each gene (E < 10−5) was selected for annotation with KOG ID. Using this KOG ID, we classified the genes into 25 KOG functional categories and then grouped them into six functional categories, namely, “information,” comprising A, J, K, and L categories of KOG; “cell cycle,” comprising B and D categories of KOG; “metabolism,” comprising C, E, F, G, H, I, P, and Q categories of KOG; “signaling and communication,” comprising O, T, and U categories of KOG; “immune and related functions,” comprising V category of KOG; and “structure and motility,” comprising M, N, W, Y, and Z categories of KOG. These six broad functional classes were defined based on the scheme suggested by Adams et al. (1) and Andrade et al. (4) (also see Refs. 37–39, 41). Two KOG categories (R, S) for “general function prediction only” and “unknown” functions were not included in the functional analysis.

Statistical Methods

Tests of significance for the differences between the proportions of genes containing dinucleotide repeats in different bins compared with global distribution of these repeats in the datasets (A, B, or C) were carried out by the binomial proportions test. The observed proportion in each bin was tested against the expected proportion, which was computed assuming no preference with respect to bin average transcriptional levels. The null hypothesis was “under neutral model,” an equi-distribution of proportion of genes with repeats is expected across all bins. Pearson’s product-moment correlation (R) was computed to examine the relationship between bin average transcription and 1) proportion of genes with repeats in different bins, 2) average intron and exon lengths of genes in bins, and 3) average GC content of genes in bins. The Pearson’s product-moment correlations (R) between average intron lengths and proportion of genes with repeats were also examined. The R statistical package (36) was used to perform the statistical tests. Partial correlations and their significance values between average transcription and repeats controlling for GC content and average intron lengths were computed with pcor and pcor.test from the ggm library of the R statistical package.

RESULTS

We used three publicly available human microarray gene expression datasets, which we designate “A” (38), “B” (44), and “C” (45). Datasets A and B were produced using HG U95A arrays, and dataset C was produced using HG U133A arrays from Affymetrix (http://www.affymetrix.com). Dataset A was obtained from blood leukocytes drawn from 13 normal human individuals (38); dataset B was obtained from 46 different human tissues, organs, and cell lines (44); and dataset C was obtained from 79 human tissues, organs, and cell lines (45). These datasets taken together offer gene expression data from both natural (from living individuals, dataset A) and artificially preserved states (from stored tissue samples, datasets B and C).

We applied two procedures to reduce or eliminate confounding effects due to random noise, since stochastic noise is an inherent property of gene expression in living systems (15). First, we considered the logarithmically transformed geometric average values of signal intensities or average difference values to represent the transcriptional level of a given gene (47). Second, we removed genes whose transcriptional levels varied randomly due to environmental causes (38). Application of the sieving procedure to any dataset will likely reduce random noise due to extrinsic environmental fluctuations (38). The resulting filtered datasets (“A”: 5,015 genes, “B”: 6,650 genes, and “C”: 11,017 genes) would be more suitable for examining the role of intrinsic factors of DNA such as the dinucleotide repeats including the most abundant (TG/CA)n=12 repeats.

Furthermore, genes were partitioned into bin sizes each of 5% of the total numbers of genes in each dataset and numbered in increasing order of bin average transcriptional levels of genes in each bin. This strategy holds the potential to unravel clear correlation patterns as opposed to using individual genes (26, 47). All three datasets compare well with respect to the transcriptional levels in all bins (dataset A and B: R = 0.93, P < 0.0001; dataset A and C: R = 0.99, P < 0.0001; dataset B and C: R = 0.97, P < 0.0001; Supplementary Fig. S1) and therefore can be used as independent resource datasets for examination of hypotheses.

Examination of the proportion of genes with uninterrupted intragenic (TG/CA)n=12 repeats in each bin and the bin average transcription revealed an inverse relationship (Fig. 2). A significant negative correlation was observed in all datasets (dataset A: R = −0.95, P < 0.0001; dataset B: R = −0.93, P < 0.0001; dataset C: R = −0.93, P < 0.0001). These results show that highly transcribed genes are poorly populated with (TG/CA)n=12 repeats. The bins with highly transcribed genes (dataset A: 15, 18–20; dataset B: 16–20; dataset C: 17–20) had lower than expected proportion of genes with (TG/CA)n=12 repeats (P < 0.05 to P < 0.0001). Conversely, bins with weakly transcribed genes (dataset A: 1–4, 6; dataset B: 1, 2, 6, 7, 9; dataset C: 1–5, 8) had higher than expected proportion of genes with (TG/CA)n=12 repeats (P < 0.04 to P < 0.0001). It is apparent that an overall agreement exists between the three datasets with respect to the general trend. It is to be noted that in the remaining bins, the differences between the expected and observed proportion of genes with (TG/CA)n=12 repeats did not show a clear statistically significant difference.

To examine whether such a pattern is specific to (TG/CA)n=12 repeats we also analyzed the proportion of other dinucleotide repeats (GA/TC)n=12, (AT)n=12, and (GC)n=12 in the three
A keeping genes with (TG/CA)\textsubscript{n} and (AT)n repeats and gene transcriptional levels. The correlation values (R) and the associated confidence values are displayed. Ensquared points: bins of genes with statistically significant lower than expected proportion of genes with repeats (P < 0.05 to P < 0.0001); encircled points: bins of genes with statistically significant higher than expected proportion of genes with repeats (P < 0.04 to P < 0.0001).

The inverse relationship was persistent, and the strength of this relationship was maintained even when the housekeeping genes from Eisenberg and Levanon (11) and from Hsiao et al. (19) were removed respectively from datasets A, B, and C. The respective correlations are: dataset A: R = −0.91, P < 0.0001; dataset B: R = −0.93, P < 0.0001; dataset C: R = −0.92, P < 0.0001.

The relationship between average intron lengths of genes and their bin average transcriptional levels displayed negative correlation (Fig. 3), reconfirming the previous observations (7) that highly transcribed genes have short introns (dataset A: R = −0.97, P < 0.0001; dataset B: R = −0.93, P < 0.0001; dataset C: R = −0.86, P < 0.0001). The average exon lengths of genes and bin average transcriptional levels in all datasets were also negatively correlated (dataset A: R = −0.9, P < 0.0001; dataset B: R = −0.95, P < 0.0001; dataset C: R = −0.87, P < 0.0001). A positive correlation was observed between average intron lengths and proportion of genes with uninterrupted intragenic (TG/CA)\textsubscript{n=12} repeats (Supplementary Fig. S2, dataset A: R = 0.94, P < 0.0001; Supplementary Fig. D3, dataset B: R = 0.90, P < 0.0001; Supplementary Fig. D4, dataset C: R = 0.96, P < 0.0001). Similarly, a positive correlation was observed between average intron lengths and proportion of genes with uninterrupted intragenic (AT)n\textsubscript{n=12} repeats (dataset A: R = 0.90, P < 0.0001; dataset B: R = 0.95, P < 0.0001; dataset C: R = 0.93, P < 0.0001) and (AT)\textsubscript{n=12} repeats (dataset A: R = 0.96, P < 0.0001; dataset B: R = 0.77, P < 0.0001; dataset C: R = 0.78, P < 0.0001).

A strong positive correlation was also observed between the average GC [%/(G+C)] content of genes in each bin and bin average transcriptional levels in all datasets (Fig. 4, dataset A: R = 0.93, P < 0.0001; dataset B: R = 0.93, P < 0.0001; dataset C: R = 0.92, P < 0.0001).

In our previous analysis, we observed that many housekeeping genes (TG/CA)\textsubscript{n=12} repeats had lower average transcriptional levels compared with those without repeats (41). Housekeeping genes are compact, highly transcribed, and poor in (TG/CA)\textsubscript{n=12} repeats (11, 41). In all datasets A, B, and C, we observed that most of the housekeeping genes were located in bins of highly transcribed genes. This pattern is in accordance with previous observations (11, 25).

To resolve whether the observed inverse relationship between transcriptional levels and incidence of (TG/CA)\textsubscript{n=12} and other dinucleotide repeats is a global phenomenon or is biased due to presence of highly transcribed housekeeping genes, we repeated the analysis after removing the housekeeping genes. The inverse relationship was persistent, and the strength of this relationship was maintained even when the housekeeping genes from Eisenberg and Levanon (11) and from Hsiao et al. (19) were removed respectively from datasets A, B, and C. The respective correlations are: dataset A: R = −0.91, P < 0.0001; dataset B: R = −0.93, P < 0.0001; dataset C: R = −0.92, P < 0.0001. The (AT)n\textsubscript{n=12} repeats also displayed similar pattern (dataset A: R = −0.88, P < 0.0001; dataset B: R = −0.91, P < 0.0001; dataset C: R = −0.88, P < 0.0001; R = −0.87, P < 0.0001). Such a persisting trend was much less evident in the case of (GA/TC)n repeats (dataset A: R = −0.50, P < 0.03 and R = −0.84, P < 0.0001; dataset B: R = −0.61, P < 0.004, R = −0.71, P < 0.0004; dataset C: R = −0.47, P < 0.04, R = −0.45, P < 0.04). It is apparent from Supplementary Tables S1, S2, and S3 that (TG/CA)\textsubscript{n=12} repeats and (AT)n\textsubscript{n=12} repeats are positively associated with (T/C)\textsubscript{n=12} repeats and negatively associated with highly expressed genes in all three datasets regardless of the breadth of transcription. However, (GA/TC)n\textsubscript{n=12} repeats showed a much weaker correlation.

The inverse relationship between the proportion of genes with uninterrupted intragenic (TG/CA)\textsubscript{n=12} repeats and gene transcriptional levels. The correlation values (R) and the associated confidence values are displayed. Ensquared points: bins of genes with statistically significant lower than expected proportion of genes with repeats (P < 0.05 to P < 0.0001); encircled points: bins of genes with statistically significant higher than expected proportion of genes with repeats (P < 0.04 to P < 0.0001).
controlling the effects of average intron lengths (most uninterrupted intragenic repeats were located in introns) and %GC content. Partial correlations for each repeat type were computed individually. Even after controlling the effects of both GC content and average intron lengths, we observed significant negative correlation of bin average transcription with (TG/CA)$_{n=12}$ repeats in datasets B and C (dataset B: pR = −0.56, P < 0.02; dataset C: pR = −0.58, P < 0.02; Supplementary Table S4). Similarly, proportion of genes with (AT)$_{n=12}$ repeats showed significant correlation with bin average transcription (dataset A: pR = −0.61, P < 0.008; dataset B: pR = −0.70, P < 0.002). The correlations of proportion of genes with (GA/TC)$_{n=12}$ repeats were not statistically significant after controlling for effects of GC content and average intron lengths.

We repeated the foregoing analysis after removing the two sets of housekeeping genes. After the housekeeping gene sets of Hsiao et al. (19) and of Eisenberg and Levanon (11), respectively, were removed, the partial negative correlations between proportion of genes with (TG/CA)$_{n=12}$ repeats and bin average transcription were persistent (dataset B: pR = −0.60, P < 0.009; dataset B: pR = −0.59, P < 0.02; dataset C: pR = −0.50, P < 0.04). Similarly, the partial negative correlation between proportion of genes with (AT)$_{n=12}$ repeats and bin average transcription were persistent (dataset A: pR = −0.60, P < 0.009; dataset B: pR = −0.50, P < 0.04; and dataset A: pR = −0.50, P < 0.04). In the case of proportion of genes with (GA/TC)$_{n=12}$ repeats the partial correlations after controlling for effects of GC and average intron lengths were not statistically significant.

**DISCUSSION**

Most (>98%) uninterrupted intragenic (TG/CA)$_{n=12}$, (AT)$_{n=12}$, and (GA/TC)$_{n=12}$ repeats were located in introns. Under the “neutral model,” the low abundance of repeats in highly transcribed genes could be due to a secondary consequence of selection either for short introns or for elevated GC composition. The observation that genes with long introns harbor more repeats compared with genes with short introns suggests a strong role for intron length as a controlling factor for abun-

---

Fig. 5. Functional composition of bins of genes with statistically significant occurrence (high or low) of (TG/CA)$_{n=12}$ repeats in dataset A. Key: (+) Bins with higher occurrence of repeats; (−) bins with lower occurrence of repeats. CC, cell cycle; IN, information; IR, immune and related functions; MET, metabolism; SC, signaling and communication; STM, structure and motility.
dance of repeats. However, we found that in case of datasets A, B, and C, 66, 61, and 56% of the longest introns (≥10 kb) of the topmost highly expressed genes (bin 19 and 20) contained no (TG/CA)$_{n\geq12}$ repeats; in contrast to the remaining groups (bin 1–18) of datasets A, B, and C, 56, 54, and 54% of the longest introns (≥10 kb) contained no (TG/CA)$_{n\geq12}$ repeats. Therefore, it appears that although long introns offer more space for accommodation of additional elements, the highly expressed genes tend to have lower proportion of (TG/CA)$_{n\geq12}$ repeats. It is apparent that the (TG/CA)$_{n\geq12}$ repeats span only a small fraction (0.095–0.145%) of the lengths of introns of genes in different bins in all three datasets. Therefore it is unlikely that the presence of these repeats would increase lengths of introns severely, thereby inflating transcriptional costs compared with large insertions such as transposons whose lengths average 300 bp (7). Therefore, selection for short introns is unlikely to be a prime mover in eliminating microsatellites, which rarely exceed 23 repeat units (46 bp) (12, 39).

In the case of (GA/TC)$_{n\geq12}$ repeats, it has been described that these repeats tend to repress transcription by stabilizing nucleosomes (9, 28). Although the trends in the case of these repeats are less strong compared with (TG/CA)$_{n\geq12}$ and (AT)$_{n\geq12}$ repeats, it appears that the (GA/TC)$_{n\geq12}$ repeats are also negatively selected for in highly expressed genes. The results of the partial correlation analysis showed that even after controlling the effects of GC and average intron length, we found that (TG/CA)$_{n\geq12}$ and (AT)$_{n\geq12}$ repeats displayed significant correlations with average transcription. In other words, although the role of repeats in the context of GC content and average intron lengths appears to be somewhat weaker, their role in transcription is clearly evident. Thus it appears that attainment of high expression is accompanied by the correlated interplay of multiple factors such as intron length, repeat content, and GC content.

Analysis of functional composition of the genes in bins which showed significant association with (TG/CA)$_{n\geq12}$ repeats (Figs. 5–7) showed that, the proportion of genes belong-
These observations suggest that (TG/CA)n repeats are negatively selected for early in the evolution of primate genomes (7). Comparison of the sequences of orthologs of ribosomal protein coding genes revealed that all human ribosomal protein coding genes are devoid of (TG/CA)n repeats. (AT)n and (GA/TC)n repeats are positively associated with genes of signaling and communication patterns, whereas they were negatively associated with the information class (39, 41). Therefore, it is probable that the low occurrence of (TG/CA)n repeats and other repeats in highly transcribed human genes is also controlled by gene function (39, 41).

The list of highly transcribed genes includes those coding for ribosomal proteins and structural proteins in accordance with previous observations reported from the analysis of human expressed sequence tags and microarray data from other non-primate genomes (7). Comparison of the sequences of orthologs of ribosomal protein coding genes revealed that all human ribosomal protein coding genes are devoid of (TG/CA)n repeats. (AT)n and (GA/TC)n repeats are positively associated with genes of signaling and communication patterns, whereas they were negatively associated with the information class (39, 41). Therefore, it is probable that the low occurrence of (TG/CA)n repeats and other repeats in highly transcribed human genes is also controlled by gene function (39, 41).


