6466–6476 *Nucleic Acids Research, 2009, Vol. 37, No. 19* *Published online 21 August 2009*
doi:10.1093/nar/gkp689

The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes

Amir B. Cohanim and Tali E. Haran*

Department of Biology, Technion, Technion City, Haifa 32000, Israel

Received July 6, 2009; Revised August 4, 2009; Accepted August 5, 2009

ABSTRACT

It is known that there are several codes residing simultaneously on the DNA double helix. The two best-characterized codes are the genetic code—the code for protein production, and the code for DNA packaging into nucleosomes. Since these codes have to coexist simultaneously on the same DNA region, both must be degenerate to allow this coexistence. *A*-tracts are homopolymeric stretches of several adjacent deoxyadenosines on one strand of the double helix, having unusual structural properties, which were shown to exclude nucleosomes and as such are instrumental in setting the translational positioning of DNA within nucleosomes. We observe, cross-kingdoms, a strong codon bias toward the avoidance of long *A*-tracts in exon regions, which enables the formation of high density of nucleosomes in these regions. Moreover, long *A*-tract avoidance is restricted exclusively to nucleosome-occupied exon regions. We show that this bias in codon usage is sufficient for enabling DNA organization within nucleosomes without constraints on the actual code for proteins. Thus, there is interdependency of the two major codes within DNA to allow their coexistence. Furthermore, we show that modulation of *A*-tract occurrences in exon versus non-exon regions may result in a unique alternation of the diameter of the '30-nm' fiber model.

INTRODUCTION

The DNA double helix carries within its base sequence multiple codes that simultaneously reside on the same DNA region (1). The best-known code is the genetic code that uses triplet of DNA codons to specify the amino-acid order of proteins (2). Another code residing in the DNA base sequence is the code for the packaging of DNA within nucleosomes. Nucleosomes frequently assume specific positions on DNA, and DNA sequence plays an essential (though not exclusive) role in establishing these preferred positions, called nucleosome positioning (3,4). To accommodate the overlapping codes within the same DNA sequence, all codes have to be degenerate to various extents, meaning that there is more than one option to specify the encoded massage.

As we all know, the triplet genetic code is degenerate, with a 64–21 mapping of codons to amino acids or stop codons, and a built in redundancy because of the possibility of wobble pairing between codons and anticodons (5). The assignments of codons to amino acids are nonrandom, and have been selected to minimize the deleterious consequences of translational reading errors (6–8) and frameshift mutations (9). In addition, synonymous codons are used with nonrandom frequencies, called codon usage bias (10,11). Differences in codon usage were shown to be correlated with the number of isoaccepting tRNA molecules, in both unicellular as well as multicellular organisms (12), optimizing the growth efficiency of cells (13). It is instructive to distinguish here between changes in the frequency of codon usage bias between genomes, which we call 'genomic codon usage', and changes within genomes (from gene-to-gene) as well as within individual genes, which we call 'local codon bias' [previously called 'major codon bias' and 'intragenic codon bias', respectively (13)]. Genomic codon usage and local codon bias have been ascribed to various biological factors, such as gene expression level (14–16), overall or local translation rate (13,17,18), gene length (19,20), protein structure (21–23), mutation rates and patterns (24,25), GC composition (18,26,27), mRNA secondary structure (28,29), or as contributing to the 10–11-bp periodicity in genomes (30,31).

The signals on DNA for the nucleosome packaging code reside in the structural properties of DNA base-pair combinations, and therefore the code is 3D and intrinsic to the conformational properties of particular DNA sequences (32,33). The conformational properties of individual base-pair steps, or the cooperative conformation of longer DNA tracts, predispose individual sequences to be in a particular conformation, or to be deformed in a particular manner upon interactions with proteins (called 'deformability') (34,35). Recent studies on

*To whom correspondence should be addressed. Tel: 972 4 8293767; Fax: 972 4 8225153; Email: bitali@tx.technion.ac.il

nucleosome positions across the whole *Saccharomyces cerevisiae* genome (36–41) and other organisms (42–44) showed that coding regions are more occupied by nucleosome relative to promoters and post-termination regions. The importance of nucleosome positioning is beyond that of merely packaging DNA passively within nucleosomes, but is important also for gene regulation. This is because nucleosome positioning controls the accessibility of regulatory proteins to specific DNA sequences (45,46). DNA positioning signals fall into two categories—local and global positioning signals. Local signals are specific dinucleotides motifs that are repeated with the helical periodicity, and they affect the local rotational and translational positioning of nucleosomes (47–50). Global positioning of nucleosomes was shown (36,51) to operate by excluding nucleosomes from DNA sequences rich in adjacent runs of homodeoxyadenosine on one strand of the double helix, called '*A*-tracts'. (DNA having a complementary structure implies that nucleosomes are excluded also by DNA sequences rich in adjacent runs of homodeoxythymidines or *T*-tracts.) *A*-tracts of length 4 bp and longer are known to switch in a cooperative manner (52,53) to a context independent structure, distinct from canonical B-DNA (54). They are characterized by bases that are inclined relative to the helix axis, propeller-twisted base pairs (with the possibility to form bifurcated hydrogen bonds), and by a narrow minor groove that becomes progressively narrower from the 5′- to 3′-end of the double helix, reaching their minimal width after 7 bp (54). These features make *A*-tracts be a dominant and conformationally rigid DNA motif that constrains B-DNA regions bordering them, especially on the 3′ side (55).

In this study, we looked at the distribution pattern of *A*-tracts (and other simple sequence motifs) in several eukaryotic genomes and show that throughout eukaryota domain codon usage in exons is biased to minimize the occurrence of long *A*-tracts. Moreover, long *A*-tract avoidance is restricted exclusively to nucleosome-occupied exon regions. This nonrandom frequency of synonymous codon usage enables the coexistence of the genetic code with the nucleosome positioning code. Based on our results we discuss possible implications for higher-order structure of chromatin.

MATERIALS AND METHODS

Data sets

The following representative organisms were studied: *S. cerevisiae* (fungi), *Arabidopsis thaliana* (plants), *Caenorhabditis elegans* (invertebrates), *Homo sapiens* (mammals) and *Danio rerio* (not-mammal vertebrates). Genome sequences and nonredundant (Refseq) genes annotations were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/refseq) on September 14, 2008. Microarray-based nucleosome map of *S. cerevisiae* was downloaded from Lee *et al.* (36). Pyrosequencing-based nucleosome map of *C. elegans* was downloaded from Johnson *et al.* (42). Annotation of secondary structure of X-ray determined protein structures from the Protein Data Bank (PDB), were downloaded from the PDBFINDER database (http://swift.cmbi.kun.nl/gv/pdbfinder). Only PDB sequences that we could locate in the NCBI Refseq data sets were used.

Control sequences

We generated three types of control sequences in which exons were regenerated by changing the codons to amino-acids assignments within each synonymous group. In control CS1 the reassignment of synonymous codons was based on natural genomic codon usage. For each studied organism, the genomic codon usage was derived from the nonredundant (Refseq) coding sequences (CDS), and from it the probability of each synonymous codon to be assigned was calculated. In this control type the local bias from genomic codon usage is cancelled, yet the natural genomic codon usage and the genetic code are maintained. In control CS2 the reassignment of synonymous codons was based on a uniform codon usage, where all synonymous codons representing an amino acid had equal probability to be assigned to that amino acid. In this control type both genomic codon usage as well as codon bias are cancelled. Uniform codon usage, with respect to a population containing an equal number of each isoaccepting tRNA molecules can be considered as random codon usage. Therefore, the only linguistic load that remains on a CS2 sequence is the genetic code. In control CS3 the reassignment of synonymous codons was based on a 'codon usage', which was created using the natural nucleotide frequency in the genome. In this control type genomic codon usage and codon bias are cancelled. This new 'codon usage' can also be considered as random codon usage. Therefore, the only linguistic load that remains on a CS3 sequence is the genetic code. We generated 100 replicates per control sequences (CS1–CS3) for each Refseq CDS sequence. The SDs retrieved from these replicates are displayed as the error-bars for CS1–CS3. In addition, we created random sequences, based on the nucleotide composition of each genome, for use in the *A*-tract alignment analyzes. For further discussion of control sequences see Supplementary Data.

Occurrence of dinucleotide repeats

We analyzed the occurrences of all possible isolated dinucleotide repeats (DNRs) of length 2–10 bp. To examine whether a DNR of a discrete length tend to be over- or under-represented in a sequence data set we compared its frequency (f_{DNR}) to the expected frequency (in the sequence data set) based on the frequency of the dinucleotide that constitutes the repeat (P_{DNR}). $\mathrm{Log}_{10}(f_{DNR}/P_{DNR})$ is defined as the 'relative frequency' of a DNR and it is used as the propensity for either abundance or avoidance of a DNR. In all figures where relative frequencies are calculated, we show the results of analysis only for DNRs that had high enough P_{DNR} to have an expected occurrence of at least one in the analyzed sequence data set and call them 'significant DNRs'.

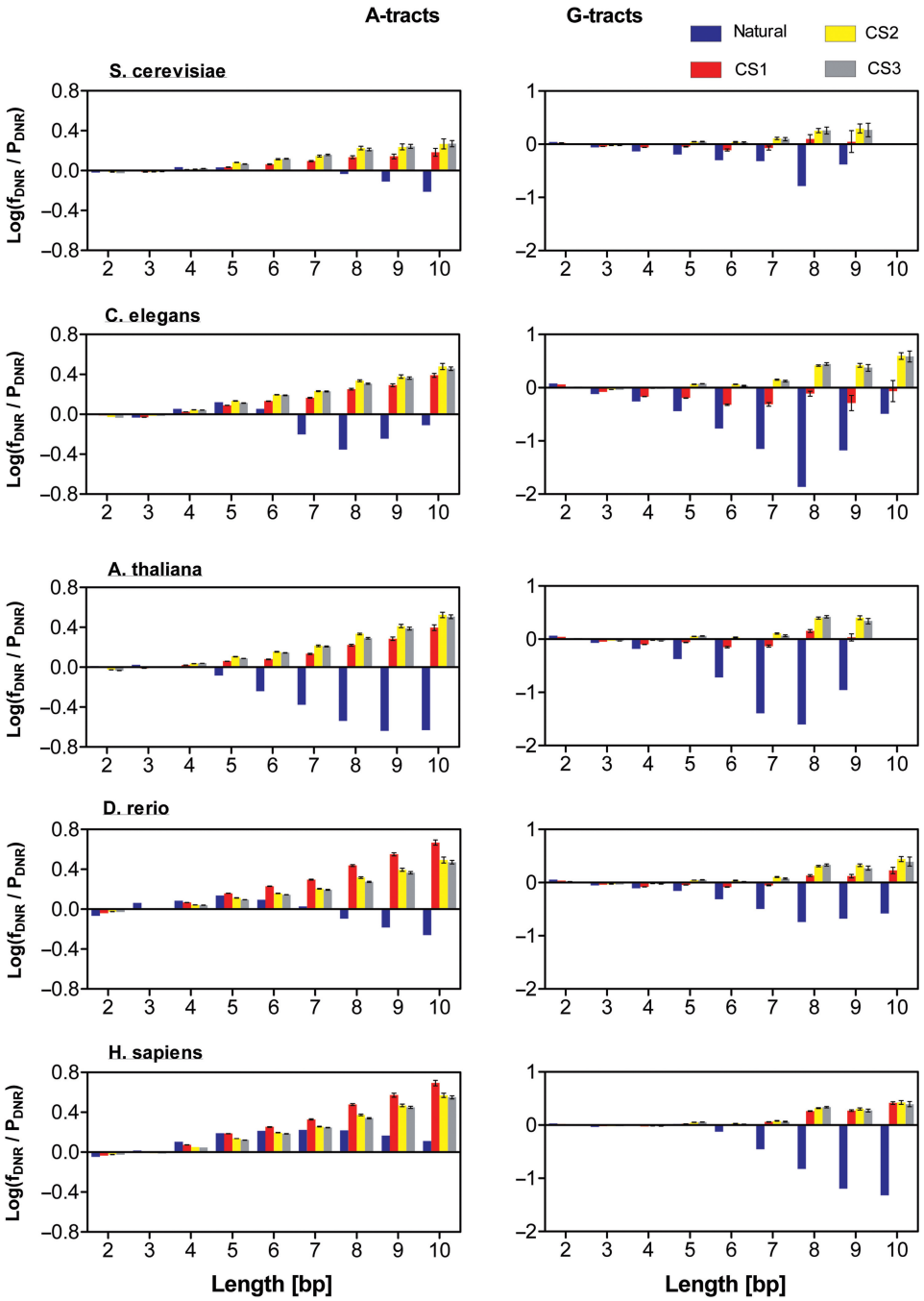

Figure 1. Relative frequencies for *A*-tracts and *G*-tracts occurrences in exons. Relative frequencies, defined as $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$, are presented for *A*-tracts and *G*-tracts in exon sequences (blue bars) and compared to controls CS1, CS2 and CS3 (red, yellow and gray bars, respectively). *X*-axis is the length (in base pair) of each tract. Only tracts with significant number of occurrences are shown.

RESULTS

Codon bias toward the avoidance of long homopolymeric tracts

We treat homopolymeric runs in this article as composed of DNRs to facilitate their comparison to hetero DNR pairs. $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$, which we call the 'relative frequency', measures the difference between the observed frequency of a DNR (f_{DNR}) and the expected frequency based on the frequency of the dinucleotide that constitutes that repeat (P_{DNR}). In Figure 1, for each organism studied here, the relative frequencies of *A*-tracts and *G*-tracts (homopolymeric runs of deoxyguanosines on one strand of the double helix) in exons are compared to controls CS1, CS2 and CS3. Each homopolymeric tract has been analyzed at increasing multiplicity, from length 2 to 10 bp. Each tract, at each length, is an isolated occurrence of such motif, thus it is not part of a longer tract built from these sequence elements.

For each organism here studied, controls CS1, CS2 and CS3 were generated by changing codons to amino-acids assignments within each synonymous group as described in 'Material and Methods' section. In Figure 1, we show that in natural exons there is a strong avoidance of $A_{n>5}$ (defined as A-tracts longer than 5 bp) and $G_{n>2}$ (defined as G-tracts longer than 2 bp) compared to controls CS1–CS3, which is common to all organisms studied here. We next analyzed the relative frequencies of all other isolated DNRs from length 2 to 10 bp. No consistent results, common to all studied organisms, were observed for any other DNR (Supplementary Figure 1). Thus, A-tracts and G-tracts are unique in this respect. The difference in relative frequencies between natural exons and the controls CS1–CS3 is always larger than 10 SD, and hence significant (SD values were obtained from the control sequences).

Avoidance of long homopolymeric tracts is achieved by local codon bias

Under constraints of preserving the genetic code avoidance of $A_{n>5}$ and $G_{n>2}$ can be achieved either by divergence of genomic codon usage from random codon usage or by local codon bias. These two approaches are explained using an example. Assume a DNA region that codes for run of three phenylalanines (FFF). Since phenylalanine codons are TTT and TTC, FFF can be coded using $A_{n>5}$ motifs: TTT-TTT-TTT, TTT-TTT-TTC or TTC-TTT-TTT, or by using non-$A_{n>5}$ motifs: TTT-TTC-TTT, TTC-TTC-TTT, TTT-TTC-TTC or TTC-TTC-TTC. Increasing the general use of codon TTC in the genome (a divergence of genomic codon usage from random codon usage) will decrease the probability for generating an $A_{n>5}$, but will also affect other DNA sequences containing codons for phenylalanine. On the other hand, local codon bias toward the avoidance of $A_{n>5}$ is a more specific approach toward the same goal: only where there is a potential for $A_{n>5}$ to be generated a local codon bias may occur. In our example, the local codon bias will be only toward the avoidance of consecutive occurrences of the codon TTT. By examining Figure 1, one could observe that the differences in relative frequency between natural exons and CS1 (differences that result from codon bias) are substantially larger than the differences in relative frequency between CS1 and CS2, or between CS1 and CS3, differences that result from the divergence of natural genomic codon usage from that of random codon usage. Thus, the avoidance of $A_{n>5}$ and $G_{n>2}$ is mostly accomplished by local codon bias. This is further supported by calculating (Supplementary Figure 2) the log ratio of the frequencies of A-tracts and G-tracts in natural exons to frequencies in CS1 [$\mathrm{Log}_{10}(f_{\mathrm{nat}}/f_{\mathrm{CS1}})$]. Relative frequencies (Figure 1) are a measure for the propensity for either abundance or avoidance of A-tracts or G-tracts, whereas $\mathrm{Log}_{10}(f_{\mathrm{nat}}/f_{\mathrm{CS1}})$ is a measure for the difference in frequency of A-tracts or G-tracts between natural exons and CS1 control sequences. As is clearly shown in Supplementary Figure 2, natural exons have substantially less $A_{n>5}$ and $G_{n>2}$.

Only $A_{n>5}$ are avoided differentially in nucleosome versus linker regions of exons

We next examined whether the observed strong codon bias toward the avoidance of $A_{n>5}$ and $G_{n>2}$ in exons is different in nucleosome versus linker regions. We used the microarray-based nucleosome map of *S. cerevisiae* (36), in which three types of nucleosome occupancy regions (or states) were determined: L-state—regions from which nucleosomes were depleted (linker regions), N-state—regions with well-positioned nucleosomes, and F-state—regions with delocalized (fuzzy) nucleosomes. Figure 2 shows the calculated relative frequencies, $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$, for A-tracts and for G-tracts separately for L- and N-state regions (both within *S. cerevisiae* exons). As in Figure 1, natural exons are compared to control sequences CS1, CS2 and CS3. A significant difference in relative frequency of $A_{n>5}$ between natural exons and controls is observed in nucleosomes (Figure 2, second panel from top), but not in linker regions (Figure 2, top). Furthermore, one could observe that, as in Figure 1, the relative frequencies of $A_{n>5}$ in nucleosomes are negligible in natural exons, and are insignificant compared to the relative frequencies in CS1 sequences. Here again, this difference in relative frequencies is substantially larger than the differences in relative frequencies between CS1 and CS2 or CS1 and CS3 sequences. Thus, there is a local codon bias toward the avoidance of $A_{n>5}$ solely in nucleosome-occupied regions. G-tracts are found to be generally avoided in exons (both in nucleosome and in linker regions), and thus do not necessarily play a role in nucleosome positioning. The same analysis was performed for *C. elegans*, using a pyrosequencing-based nucleosome map (42). Since only part of the pyrosequences were located in the genome, and no clear annotation of linker regions was given by the authors, we considered as linker regions the first 20 bp flanking nucleosomes from each side. Even with such a rough definition of linker regions we observe that only in nucleosome-occupied exon regions there is a significant and consistent avoidance of $A_{n>5}$ (Supplementary Figure 3).

The genetic code itself is unconstrained by $A_{n>5}$ avoidance

We next examined whether the genetic code itself is constrained to facilitate the observed avoidance of $A_{n>5}$ in exon regions occupied by nucleosomes. Using the example given above, such constrain would be the avoidance of phenylalanine runs (FF...) in amino-acid regions corresponding to exon regions occupied by nucleosomes. As discussed above, the only linguistic load that remains on CS2 and CS3 sequences is the genetic code. Therefore, the only way to have a difference in the relative frequency of $A_{n>5}$ between CS2 (or CS3) sequences in nucleosomes and in linkers is if the amino-acid sequences themselves are biased. However, the relative frequencies of $A_{n>5}$ in CS2 (or CS3) sequences are always the same in nucleosome and in linker regions within error (Figure 2). Hence, no constraint is imposed on the amino-acid sequences themselves to facilitate the specific distribution of $A_{n>5}$ (i.e. avoidance in nucleosome regions). Furthermore, a comparison between DNA

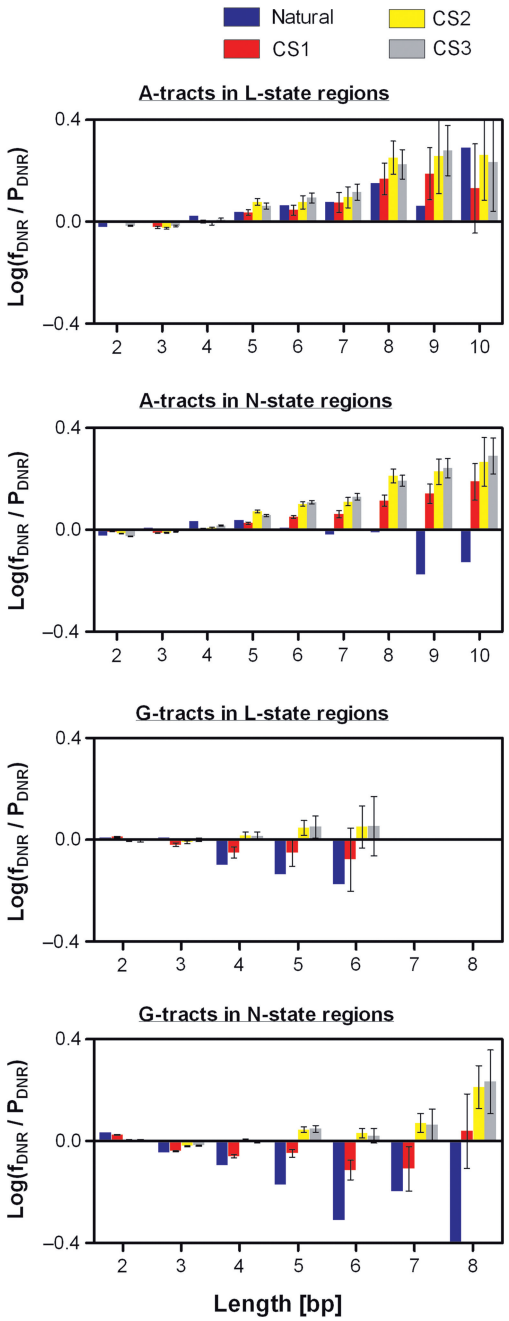

Figure 2. $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$ for *A*-tracts and *G*-tracts presented separately for linker (*L*-state) regions and nucleosome (*N*-state) regions within exons. Exon sequences (blue bars) are compared to controls CS1, CS2 and CS3 (red, yellow and gray bars respectively). *X*-axis is the length (in base pair) of each tract. Only tracts with significant number of occurrences are shown.

regions coding for the two main groups of protein secondary structures: helix (3_{10} helix, alpha helix and pi helix) and strand (beta sheet and beta bridge) shows no significant differences in the relative frequencies of $A_{n>5}$ between these two groups (Figure 3).

Exons: islands of *A*-tract dearth

We examined *A*-tract occurrences in the various vicinities within coding regions, in the five representative eukaryotic genomes studied here, by comparing the frequencies of $A_{n>5}$ within exons to those in all adjacent regions that border them—(i) upstream to start codon; (ii) introns; (iii) post stop-codon regions. We use in this analysis only nonexon regions that are free from an overlapping code for proteins. In Figure 4, protein coding genes were aligned with respect to the start codon (column A), exon–intron junction (column B), intron–exon junction (column C) or the stop codon (column D). The frequencies of occurrence of nucleotides A and T (f_{A} and f_{T}, respectively) coming only from $A_{n>5}$ are displayed for natural sequences, and for two control based on nucleotide frequency in the genomes (Figure 4). The first control is random sequences generated based on the genome nucleotide composition, and the second control is CS3 control sequences for exons. In all alignments, across all organisms, the same phenomenon is observed: all regions flanking exons are much more abundant with $A_{n>5}$. These differences in $A_{n>5}$ frequency, between exons and nonexon regions, persist throughout the aligned sequences, whether the analysis is carried out up to 1000 bp from the exon/nonexon junction (Figure 4), or carried out for up to 10 000 bp into nonexon regions (Supplementary Figure 4). For each individual panel in Figure 4 we used a two-tailed Mann–Whitney U-test (56) to examine whether the median frequency of $A_{n>5}$ in exon regions is significantly different from adjacent nonexons regions. The result is that across all organisms, these differences are significant with $P<0.0001$. To avoid a bias from junction areas (start codon, exon–intron junction, intron–exon junction and stop codon) we excluded from the statistical test the first 10 bp prior and post to each junction. The differences in $A_{n>5}$ frequency between exons and nonexons regions are not only significant on average, but are also distinctive in the junction regions. Looking at the alignment graphs (Figure 4), we notice that there is a sharp drop in $A_{n>5}$ frequencies when entering to exon regions, and a sharp rise in $A_{n>5}$ frequencies when exiting exon regions. We carried Spearman correlations between all pair-wised combinations of blue lines displayed in Figure 4. The absolute values of the correlation coefficients ranged from 0.65 to 0.89, all with $P<0.0001$. Thus, there is a similar segmentation of $A_{n>5}$ frequencies relative to any junction (i.e. high $A_{n>5}$ frequencies in nonexon regions relative to exons), whether entering or exiting exon regions, and it is common to all of the organisms studied here.

$A_{n>5}$ deficiency in exons does not result from regional differences in DNA composition

To exclude the possibility that the differences in $A_{n>5}$ frequencies between exons and nonexons are a side effect resulting from regional differences in nucleotide and dinucleotide compositions, we separately calculated the relative frequencies, $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$, of *A*-tracts in exons and in regions flanking them (Supplementary Figure 5). To this end, we divided the nonredundant gene data set of each organism to 10 independent groups from which we calculated the average and error-bars of relative frequencies (Supplementary Figure 5).

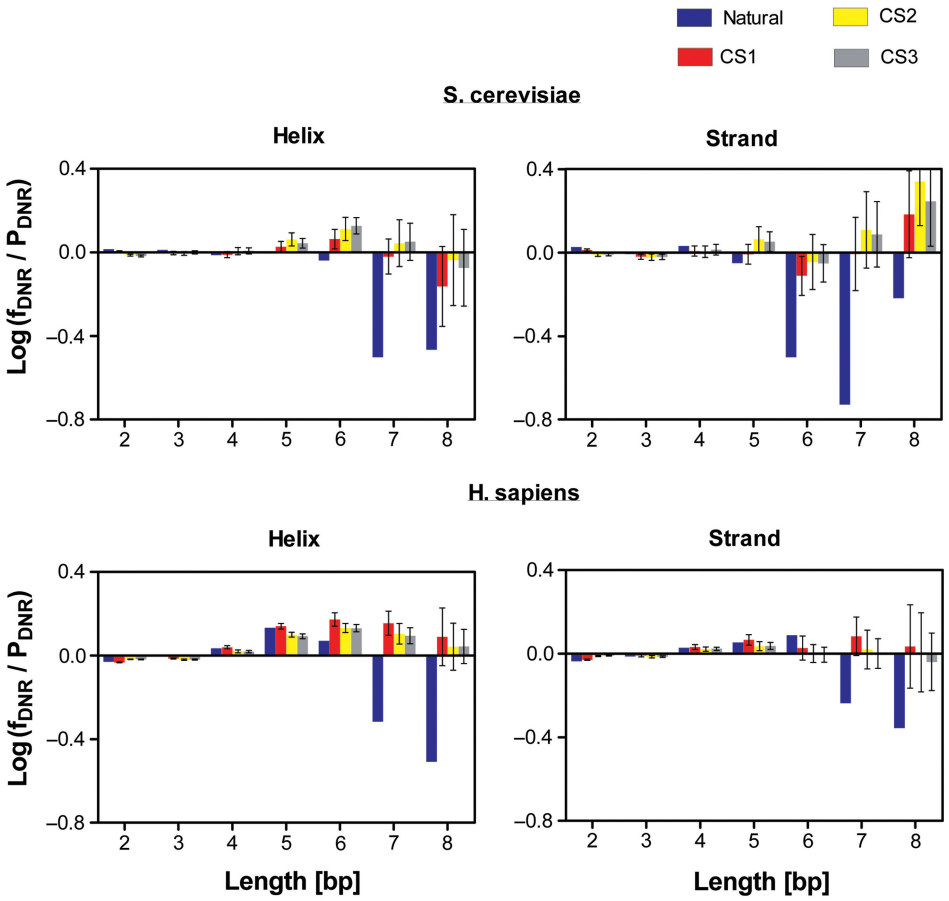

Figure 3. $\mathrm{Log}_{10}(f_{\mathrm{DNR}}/P_{\mathrm{DNR}})$ for *A*-tracts presented separately for regions coding for helices (3_{10} helix, alpha helix and pi helix) and strands (beta sheet and beta bridge). Exon sequences (blue bars) are compared with controls CS1, CS2 and CS3 (red, yellow and gray bars respectively). *X*-axis is the length (in base pair) of the *A*-tracts. Shown are results for significant DNRs only.

In all organisms, the relative frequencies were much higher in regions flanking the exons than within exons. Since the groups were independent groups, we were able to use a simple binomial test for significance: we counted as success the number of groups (equivalent to number of trials) in which the relative frequency in a 'flanking region' was above that observed in exons. The test was separately done for each flanking region and for each *A*-tract longer than 5 bp. In all cases we got 10 out of 10 successes. Thus, for all *A*-tracts longer than 5 bp the relative frequency in the flanking regions is significantly higher than in exons ($P = 0.00098$).

$A_{n>5}$ deficiency in all organisms is relative to the start codon

To make sure that the drop in $A_{n>5}$ frequency is indeed relative to the start codon, and not the transcription start site (TSS), we aligned the protein coding genes with respect to the start codon, as before, but here we included in the analysis only regions that were downstream to the TSS (Supplementary Figure 6). As before, we made sure that nonexon regions were free from overlapping protein-coding signals. Because the number of nonexon regions conforming to the constraint listed above are much smaller than the overall nonredundant genes available, we display in Supplementary Figure 6 the frequencies averaged between organisms, thus increasing the signal to noise ratio. A clear difference in the frequency of $A_{n>5}$ between upstream and downstream regions (relative to start codon) is observed. We carried a two-tailed Mann–Whitney U-test (56), separately for each organism, to examine whether the median frequency of $A_{n>5}$ in exon regions is significantly different from regions upstream to start codon. The result is that across all organisms, the differences are significant with $P < 0.0001$. Thus, the sharp drop in $A_{n>5}$ frequency between promoters and exons is indeed relative to start codons.

$A_{n>5}$ frequency is anti-corrleated with nucleosome occupancy

We combined the well-positioned nucleosome state (*N*-state) with the delocalized nucleosome state (*F*-state) of protein coding genes of *S. cerevisiae* and *C. elegans* to enhance the pattern of nucleosome occupancy. In Figure 5, the frequency of *N*-state + *F*-state was calculated for each position along protein coding genes, which were aligned as in Figure 4. Since *S. cerevisiae* does

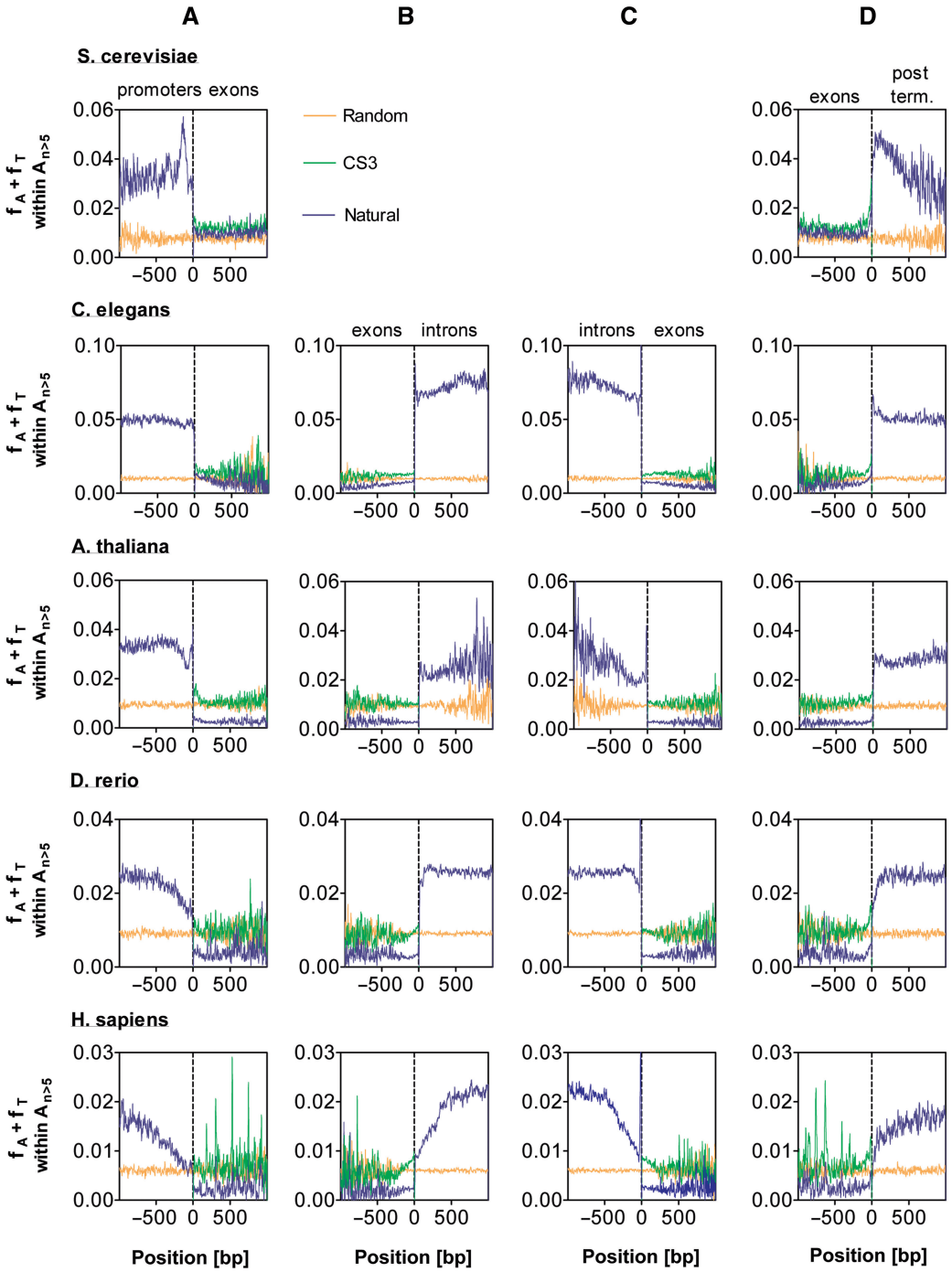

Figure 4. Alignments of $A_{n>5}$ with respect to exon/nonexon junctions. Protein coding genes are aligned with respect to start codon (column **A**), exon–intron junction (column **B**), intron–exon junction (column **C**), or stop codon (column **D**). Only nonexon regions free from overlapping genetic code were used. The frequencies of occurrence of nucleotides A and T (f_A and f_T, respectively) coming only from $A_{n>5}$ are displayed for natural sequences (blue lines), random sequences generated based on genome nucleotide composition (orange lines), and CS3 control sequences for exons (green lines). Since *S. cerevisiae* does not have sufficient number of introns for analysis, only alignments relative to start and stop codons are presented for *S. cerevisiae*.

not have sufficient number of introns for analysis, only alignments relative to start and stop codons are presented for *S. cerevisiae*. The same analysis was carried out for longer-range display (up to 10 000 bp of nonexon regions) of *C. elegans* alignments (Supplementary Figure 7). A two-tailed Mann–Whitney U-test (56) was carried out to examine whether the median frequency of *N*-state + *F*-state in exon regions is significantly different from adjacent nonexons regions. The result is that for both *S. cerevisiae* and *C. elegans*, in all regions shown in Figure 5, the differences are significant with $P < 0.0001$. Thus, exons are significantly more occupied by

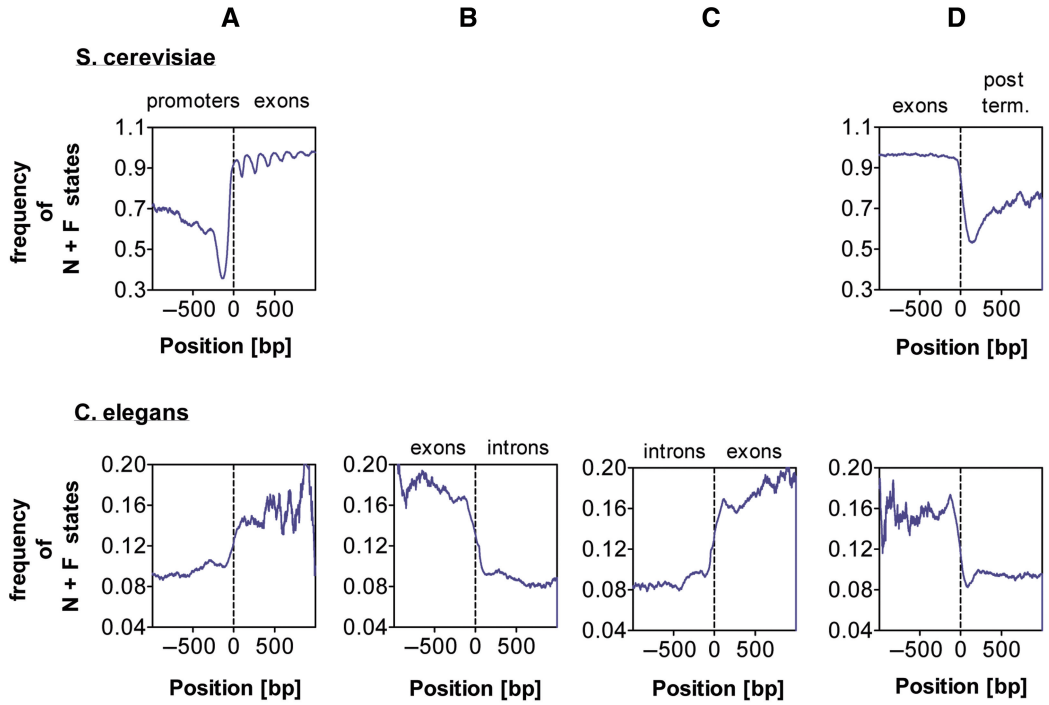

Figure 5. Alignments of nucleosome occupancy with respect to exon/nonexon junctions. Protein coding genes are aligned with respect to start codon (column **A**), exon–intron junction (column **B**), intron–exon junction (column **C**), or stop codon (column **D**). Only nonexon regions free from overlapping genetic code were used here. The frequency of *N*-state (well-positioned nucleosome) + *F*-state (delocalized nucleosome state) is calculated for each position along the aligned sequences of *S. cerevisiae* and *C. elegans*. Since *S. cerevisiae* does not have sufficient number of introns for analysis, only alignments relative to start and stop codons are presented for *S. cerevisiae*.

nucleosomes relative to all flanking regions, including introns. When we compare Figure 4 ($A_{n>5}$ frequency) to Figure 5 (nucleosome occupancy) a clear anti-correlation is observed for all regions. We carried out Spearman correlations between natural $A_{n>5}$ frequency graphs (Figure 4) and nucleosome occupancy graphs (Figure 5) of the same regional alignment (e.g. blue line in the top left panel of Figure 4 versus the blue line in panel A of Figure 5). The correlation coefficients ranged from –0.67 to –0.92, all with $P < 0.0001$.

Linker lengths are different between coding and noncoding regions

Since, as shown in Figure 5, exons are more densely occupied by nucleosomes relative to all regions bordering them, linker regions in exons should be by default shorter than linkers in regions flanking exons. The average lengths of linker regions (or nucleosome-depleted regions) of protein coding genes were aligned relative to the start codon (Figure 6A) or the stop codon (Figure 6B) of *S. cerevisiae*. After alignment each position along the sequence (*x*-axis) represents the linker size calculated by averaging the lengths of all linkers covering that position. From this analysis we determine that the average size of linkers in exons is 51 ± 0.3 bp, whereas for regions upstream to the start codon and downstream to the stop codon the average size is 264 ± 0.9 bp and 220 ± 0.6 bp, respectively. In Figure 6 we used only *L*-state regions as linkers. However, linkers could possibly exist also within *F*-state regions, which are especially abundant in exons (36) and which we did not include in our analysis. DNA regions are more likely to be denoted as fuzzy when they are occupied by closely packed nucleosomes (37), thus the linker size in exons may be even smaller than the value observed here, 51 ± 0.3 bp. The observation that exons have shorter linkers compared to regions bordering them can suggest that they may be used as a structural distinction of exons at the level of the 30-nm fiber. A simplified schematic model for such a pattern is shown in Figure 7 and discussed below.

DISCUSSION

The assignment of codons to amino acids is not random, and was selected to minimize the impact of translational errors (6–8), and frameshift mutations (9). These nonrandom assignments were recently shown to be optimal for carrying additional messages in DNA sequences, on top of the genetic message (9). Here, we show another aspect of nonrandomness in the genetic code, which allow the genetic code to coexist with the nucleosome positioning code. This is the nonrandom frequency in which synonymous codons are used. The genetic code is essentially universal to all living organisms, and thus has been probably established very early in the history of life. Hence, it may have predated the conversion of genomes as condensed naked DNA, such as in viruses, to genomes in eukaryotic cells, in which DNA is packaged within nucleosomes. This

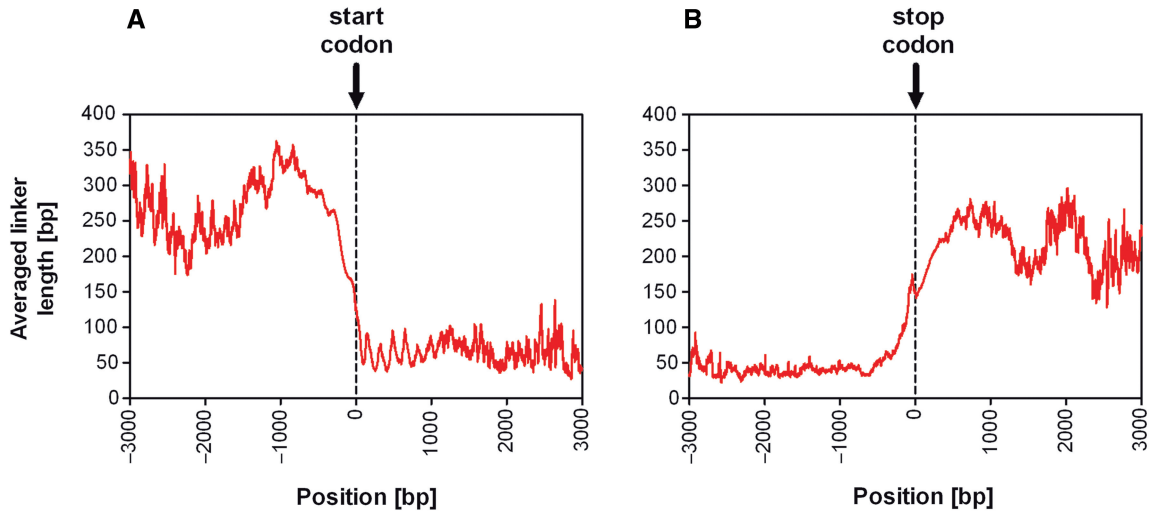

Figure 6. Alignment of linker lengths with respect to start codon or stop codon. *Saccharomyces cerevisiae* protein coding genes, aligned with respect to start codon (**A**) or stop codon (**B**), are presented. Only nonexon regions free from overlapping genetic code were used. Each position along the sequence represents the linker size calculated by averaging the lengths of all linkers covering that position.

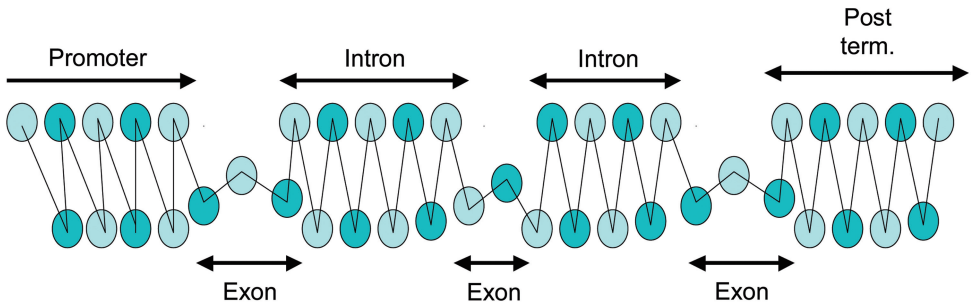

Figure 7. Schematic illustration of structural modulation of the 30-nm fiber along a eukaryotic gene. In this model the diameter and packing density of the 30-nm fiber vary along the gene due to changes in the lengths of DNA linker regions. Circles represent the nucleosome core particles; lines represent DNA linkers (or nucleosome free regions).

'chance' occurrence of the availability of a redundant code has facilitated the 'necessity' to package the several orders of magnitude more DNA inside eukaryotic cells, with the only new tinkering being different frequency of codon usage. Local codon bias enables the establishment of regions with low $A_{n>5}$ density, and consequentially regions with high nucleosome occupancy.

Previous studies showed generally that relative to coding regions, promoters are significantly richer with *A*-tracts and less occupied by nucleosomes (36,37,41). However, to our knowledge no detailed analysis has been carried to quantitatively examine *A*-tract occurrences and its relationship to nucleosome occupancy in the various vicinities within coding regions. In particular, no explicit quantitative analysis regarding nucleosome occupancies in intron sequences was carried out. Our results show that exons are significantly more occupied by nucleosomes relative to all flanking regions, including introns.

Our results show that local codon bias is used to reduce the occurrence of $A_{n>5}$ tracts in exons, and that this is related to nucleosome positioning in exons. However, several other explanations could be given to account for the avoidance of long *A*-tracts in exons. First, it is known that homopolymeric DNA tracts, and other simple sequence repeats (SSRs), can facilitate the slippage of DNA polymerase during replication, which will expand and elongate the repeat (57). It has been suggested (58) that homopolymeric sequences are over-represented in noncoding sequences, relative to coding sequences, because coding sequences are subjected to stronger selection and hence would not allow the accumulation of long *A*-tracts via slippage mechanism. This idea is contradictory to the observations in this study, where $A_{n>5}$ are avoided in exons only in regions occupied by nucleosomes, and not in nucleosome free regions (linkers) of exons. In addition, polymerase slippage during replication, generating long SSRs or microsatellites, is known to occur also for hetero SSRs, such as AT islands, and not only for homopolymeric repeats (59–61), whereas our observations are limited exclusively to homopolymeric *A*-tract sequences.

Furthermore, analysis of microsatellite distribution in different genomes suggests that strand slippage alone is not sufficient to explain the distribution patterns of microsatellites (62).

A different potential cause for the avoidance of long *A*-tract in exons is the structure of the minor groove in *A*-tracts. As mentioned above, the minor groove of *A*-tracts narrows asymmetrically in the 5′- to 3′-direction and reaches its narrowest width at length 7 bp (54). This narrowing of the minor groove was shown to cause termination of DNA synthesis by reverse transcriptase in retroviruses (63). However, this suggestion as well is not consistent with our observations that $A_{n>5}$ are avoided in exons only when the region is occupied by nucleosomes, and not in exonic linker regions. Thus, we must conclude that $A_{n>5}$ avoidance and occurrence patterns are primarily signals for nucleosome positioning and nucleosome free regions, respectively.

The nucleosome core structure represents the first level of DNA compaction. The next level is called the 30-nm fiber (64). The two main models that describe the 30-nm fiber structure are called the solenoid model and the cross-linker model (65). There is now compelling experimental evidence indicating that the 30-nm fiber can adopt the organization proposed by the cross-linker model (65). The electron microscopy (EM) study of Robinson *et al.* (66) showed that DNA linker length positively correlates with the diameter and packing density of the 30-nm fiber in the cross-linker model. As the linker length increases, the diameter of the 30-nm fiber increases. This reduces steric exclusion effects in nucleosome-to-nucleosome orientation, and allows a more dense packing along the DNA. Our model (Figure 7) amalgamates our observations with those from the EM study on the relationship between DNA linker length, packing density and diameter in the 30-nm structure. We speculate that this modulation of packing density and diameter size can be used to distinguish exon from nonexon regions at the level of the 30-nm fiber, which may perhaps be of use in gene regulation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The Israel Science Foundation. Funding for open access charge: funds for the promotion of research in the Technion.

Conflict of interest statement. None declared.

REFERENCES

1. Trifonov,E.N. (1989) The multiple codes of nucleotide sequences. *Bull. Math. Biol.*, **51**, 417–432.
2. Nirenberg,M., Caskey,T., Marshall,R., Brimacombe,R., Kellogg,D., Doctor,B., Hatfield,D., Levin,J., Rottman,F., Pestka,S. *et al.* (1966) The RNA code and protein synthesis. *Cold Spring Harb. Symp. Quant. Biol.*, Vol. 31. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., pp. 11–24.
3. Rando,O.J. and Ahmad,K. (2007) Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.*, **19**, 250–256.
4. Schnitzler,G.R. (2008) Control of nucleosome positions by DNA sequence and remodeling machines. *Cell Biochem. Biophys.*, **51**, 67–80.
5. Crick,F.H. (1966) Codon—anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19**, 548–555.
6. Woese,C.R. (1965) On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA*, **54**, 1546–1552.
7. Haig,D. and Hurst,L.D. (1991) A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, **33**, 412–417.
8. Freeland,S.J. and Hurst,L.D. (1998) The genetic code is one in a million. *J. Mol. Evol.*, **47**, 238–248.
9. Itzkovitz,S. and Alon,U. (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.*, **17**, 405–412.
10. Grantham,R., Gautier,C., Gouy,M., Mercier,R. and Pave,A. (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.*, **8**, r49–r62.
11. Sharp,P.M., Cowe,E., Higgins,D.G., Shields,D.C., Wolfe,K.H. and Wright,F. (1988) Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens: a review of the considerable within-species diversity. *Nucleic Acids Res.*, **16**, 8207–8211.
12. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.
13. Kurland,C.G. (1991) Codon bias and gene expression. *FEBS Lett.*, **285**, 165–169.
14. Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, **9**, r43–r74.
15. Sharp,P.M., Tuohy,T.M. and Mosurski,K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.*, **14**, 5125–5143.
16. Holm,L. (1986) Codon usage and gene expression. *Nucleic Acids Res.*, **14**, 3075–3087.
17. Xia,X. (1996) Maximizing transcription efficiency causes codon usage bias. *Genetics*, **144**, 1309–1320.
18. Kanaya,S., Yamada,Y., Kinouchi,M., Kudo,Y. and Ikemura,T. (2001) Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.*, **53**, 290–298.
19. Eyre-Walker,A. (1996) Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? *Mol. Biol. Evol.*, **13**, 864–872.
20. Duret,L. and Mouchiroud,D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl Acad. Sci. USA*, **96**, 4482–4487.
21. Adzhubei,A.A., Adzhubei,I.A., Krasheninnikov,I.A. and Neidle,S. (1996) Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett.*, **399**, 78–82.
22. Xie,T. and Ding,D. (1998) The relationship between synonymous codon usage and protein structure. *FEBS Lett.*, **434**, 93–96.
23. Gupta,S.K., Majumdar,S., Bhattacharya,T.K. and Ghosh,T.C. (2000) Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.*, **269**, 692–696.
24. Bulmer,M. (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics*, **129**, 897–907.
25. Sueoka,N. (1992) Directional mutation pressure, selective constraints, and genetic equilibria. *J. Mol. Evol.*, **34**, 95–114.
26. Sueoka,N. and Kawanishi,Y. (2000) DNA G+C content of the third codon position and codon usage biases of human genes. *Gene*, **261**, 53–62.
27. Knight,R.D., Freeland,S.J. and Landweber,L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.*, **2**, research0010.0011–0010.0013.

28. Zama,M. (1990) Codon usage and secondary structure of mRNA. *Nucleic Acids Symp. Ser.*, **22**, 93–94.
29. Gambari,R., Nastruzzi,C. and Barbieri,R. (1990) Codon usage and secondary structure of the rabbit alpha-globin mRNA: a hypothesis. *Biomed. Biochim. Acta*, **49**, S88–S93.
30. Cohanim,A.B., Kashi,Y. and Trifonov,E.N. (2006) Three sequence rules for chromatin. *J. Biomol. Struct. Dyn.*, **23**, 559–566.
31. Cohanim,A.B., Trifonov,E.N. and Kashi,Y. (2006) Specific selection pressure at the third codon positions: contribution to 10- to 11-base periodicity in prokaryotic genomes. *J. Mol. Evol.*, **63**, 393–400.
32. Travers,A.A. and Klug,A. (1990) In Cozzarelli,N.R. and Wang,J.C. (eds), *DNA Topology and Its Biological Effects*. Cold Spring Harbor Press, Cold Spring Harbor, pp. 57–106.
33. Travers,A. and Drew,H. (1997) DNA recognition and nucleosome organization. *Biopolymers*, **44**, 423–433.
34. Travers,A.A. (1995) In Lilley,D.M.J. (ed.), *DNA-Protein: Structural interactions*. IRL Press, Oxford, pp. 49–75.
35. Luisi,B. (1995) In Lilley,D.M.J. (ed.), *DNA-Protein: Structural interactions*. IRL Press, Oxford, pp. 1–48.
36. Lee,W., Tillo,D., Bray,N., Morse,R.H., Davis,R.W., Hughes,T.R. and Nislow,C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.
37. Yuan,G.C., Liu,Y.J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, **309**, 626–630.
38. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thastrom,A., Field,Y., Moore,I.K., Wang,J.P. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
39. Albert,I., Mavrich,T.N., Tomsho,L.P., Qi,J., Zanton,S.J., Schuster,S.C. and Pugh,B.F. (2007) Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature*, **446**, 572–576.
40. Song,J.S., Liu,X., Liu,X.S. and He,X. (2008) A high-resolution map of nucleosome positioning on a fission yeast centromere. *Genome Res.*, **18**, 1064–1072.
41. Field,Y., Kaplan,N., Fondufe-Mittendorf,Y., Moore,I.K., Sharon,E., Lubling,Y., Widom,J. and Segal,E. (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.*, **4**, e1000216.
42. Johnson,S.M., Tan,F.J., McCullough,H.L., Riordan,D.P. and Fire,A.Z. (2006) Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.*, **16**, 1505–1516.
43. Ozsolak,F., Song,J.S., Liu,X.S. and Fisher,D.E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.
44. Mavrich,T.N., Jiang,C., Ioshikhes,I.P., Li,X., Venters,B.J., Zanton,S.J., Tomsho,L.P., Qi,J., Glaser,R.L., Schuster,S.C. *et al.* (2008) Nucleosome organization in the Drosophila genome. *Nature*, **453**, 358–362.
45. Lu,Q., Wallrath,L.L. and Elgin,S.C. (1994) Nucleosome positioning and gene regulation. *J. Cell Biochem.*, **55**, 83–92.
46. Wolffe,A.P. (1994) Nucleosome positioning and modification: chromatin structures that potentiate transcription. *Trends Biochem. Sci.*, **19**, 240–244.
47. Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
48. Ioshikhes,I., Bolshoy,A., Derenshteyn,K., Borodovsky,M. and Trifonov,E.N. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.
49. Lowary,P.T. and Widom,J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.
50. Widlund,H.R., Cao,H., Simonsson,S., Magnusson,E., Simonsson,T., Nielsen,P.E., Kahn,J.D., Crothers,D.M. and Kubista,M. (1997) Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.*, **267**, 807–817.
51. Mavrich,T.N., Ioshikhes,I.P., Venters,B.J., Jiang,C., Tomsho,L.P., Qi,J., Schuster,S.C., Albert,I. and Pugh,B.F. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.*, **18**, 1073–1083.
52. Haran,T.E. and Crothers,D.M. (1989) Cooperativity in A-tract structure and bending properties of composite TnAn blocks. *Biochemistry*, **28**, 2763–2767.
53. Nadeau,J.G. and Crothers,D.M. (1989) Structural basis for DNA bending. *Proc. Natl Acad. Sci. USA*, **86**, 2622–2626.
54. Haran,T.E. and Mohanty,U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Quart. Rev. Biophys.*, **42**, 41–81.
55. Merling,A., Sagaydakova,N. and Haran,T.E. (2003) A-tract polarity dominate the curvature in flanking sequences. *Biochemistry*, **42**, 4978–4984.
56. Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
57. Levinson,G. and Gutman,G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.*, **4**, 203–221.
58. Dechering,K.J., Cuelenaere,K., Konings,R.N. and Leunissen,J.A. (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.*, **26**, 4056–4062.
59. Debrauwere,H., Gendrel,C.G., Lechat,S. and Dutreix,M. (1997) Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites. *Biochimie*, **79**, 577–586.
60. Bois,P. and Jeffreys,A.J. (1999) Minisatellite instability and germline mutation. *Cell Mol. Life Sci.*, **55**, 1636–1648.
61. Jackson,J.A., Trevino,A.V., Herzig,M.C., Herman,T.S. and Woynarowski,J.M. (2003) Matrix attachment region (MAR) properties and abnormal expansion of AT island minisatellites in FRA16B fragile sites in leukemic CEM cells. *Nucleic Acids Res.*, **31**, 6354–6364.
62. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
63. Lavigne,M. and Buc,H. (1999) Compression of the DNA minor groove is responsible for termination of DNA synthesis by HIV-1 reverse transcriptase. *J. Mol. Biol.*, **285**, 977–995.
64. van Holde,K. and Zlatanova,J. (2007) Chromatin fiber structure: Where is the problem now? *Semin. Cell Dev. Biol.*, **18**, 651–658.
65. Wu,C., Bassett,A. and Travers,A. (2007) A variable topology for the 30-nm chromatin fibre. *EMBO Rep.*, **8**, 1129–1134.
66. Robinson,P.J., Fairall,L., Huynh,V.A. and Rhodes,D. (2006) EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc. Natl Acad. Sci. USA*, **103**, 6506–6511.