BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(9)

3-base periodicity in coding DNA is affected by intercodon dinucleotides

Joaquín Sánchez

Facultad de Medicina, Universidad Autónoma del Estado de Morelos, Cuernavaca, 62020 México; Email: joaquin.sanchez@microbio.gu.se; Phone: 52-777-3184797

Received July 07, 2011; Accepted July 12, 2011; Published July 19, 2011

Abstract:
All coding DNAs exhibit 3-base periodicity (TBP), which may be defined as the tendency of nucleotides and higher order n-tuples, e.g. trinucleotides (triplets), to be preferentially spaced by 3, 6, 9 etc, bases, and we have proposed an association between TBP and clustering of same-phase triplets. We here investigated if TBP was affected by intercodon dinucleotide tendencies and whether clustering of same-phase triplets was involved. Under constant protein sequence intercodon dinucleotide frequencies depend on the distribution of synonymous codons. So, possible effects were revealed by randomly exchanging synonymous codons without altering protein sequences to subsequently document changes in TBP via frequency distribution of distances (FDD) of DNA triplets. A tripartite positive correlation was found between intercodon dinucleotide frequencies, clustering of same-phase triplets and TBP. So, intercodon C|A (where "|" indicates the boundary between codons) was more frequent in native human DNA than in the codon-shuffled sequences; higher C|A frequency occurred along with more frequent clustering of C|AN triplets (where N jointly represents A, C, G and T) and with intense CAN TBP. The opposite was found for C|G, which was less frequent in native than in shuffled sequences; lower C|G frequency occurred together with reduced clustering of C|GN triplets and with less intense CGN TBP. We hence propose that intercodon dinucleotides affect TBP via same-phase triplet clustering. A possible biological relevance of our findings is briefly discussed.

Keywords: Period-3 DNA structure; neighboring codon choice; dinucleotide frequency; human coding DNA; codon junctions; same-phase triplet clustering; frequency distribution of distances.

Background:
The 3-base periodicity (TBP) is an intrinsic property of all coding DNA **[1-6]** that is characterized by the disposition of nucleotides and higher order n-tuples, e.g. trinucleotides (triplets), so that they are preferentially separated by multiples of 3 bases, i.e. 3, 6, 9, 12 etc. TBP, or period-3 structure, has been found to be present in exons but not in intron sequences **[1]**. TBP can be used to identify coding regions in genomes **[4-6]** and the length of the period equal to three has been proposed to be caused by the usage bias of nucleotides within synonymous codons **[7]**. However, we have put forward a model wherein TBP was proposed to be produced by clustering of same-phase triplets **[8]**. A role for triplet clusters in TBP has also been suggested by others **[7]**. Nonetheless, it has not yet been established what causes clustering of same-phase triplets and therefore TBP; however, two levels of influence are possible: clustering associated to protein sequence and clustering independent of protein sequence. At the protein level same-phase triplet clustering and hence TBP occur because in natural protein-coding sequences codons are never used in equal proportions; that is, the influence of protein is reduced to codon composition. This would explain why computer generated sequences with the same codon usage as native sequences expressed TBP **[3]** and why TBP persisted after shuffling of codons **[9]**, although differences in TBP were pointed out **[9]**. We then postulate that codon-composition-independent differences in TBP are due to protein-sequence-independent intercodon dinucleotide tendencies. We here aim at determining if such hypothesis is correct and whether intercodon tendencies reflect in TBP and in the clustering of same-phase triplets.

Due to the so-called degeneracy of the genetic code with the exception of two cases (methionine and tryptophan) the cell can use more than one synonymous codon to specify the same amino acid in proteins. Nevertheless, in most coding DNA, especially in higher organisms such as vertebrates, synonymous codons are not randomly distributed **[10]**. The existence of this bias in the distribution of synonymous codons, also called neighboring codon choice, has motivated comprehensive analyses of dicodon frequencies **[11]**. Why such bias exists is not yet clear, but it may be associated to local RNA secondary structure **[12]**, to protein translation **[13]** or it may have a role in protein folding **[14]**. Independently of its origin, the non random distribution of synonymous codons reflects in intercodon dinucleotide frequencies **[15]**. Then, to investigate if and how intercodon dinucleotides affected TBP, we disrupted the natural distribution of synonymous codon by randomly shuffling them without changing protein sequences. We found differences between native and synonymous-codon-shuffled sequences in TBP as well as in clustering of same-phase triplets; we consequently propose that intercodon dinucleotides affect TBP via changes in same-phase triplet clustering.

Methodology:
The human ORFeome version hORFeome v3.1 consisting of over 12,000 human coding sequences was downloaded from http://horfdb.dfci.harvard.edu/. Annotations were edited out and sequences were subsequently semi-manually curated to eliminate: out-of-frame sequences, sequences lacking stop codons,

and sequences not starting with ATG. One thousand randomly selected coding sequences were analyzed. Other used sequences were obtained from ncbi.nlm.nih.gov and equally curated before analysis. Sequences were collected and analyzed using either one or a combination of the programs Word (Microsoft Inc, MS), Excel (MS), OMIGA 2.0 (Oxford Molecular Ltd, UK), GSCalc 6.0 (http://www.jps-development.com) or the program SWAAP **[16]**. For codon randomization each coding sequence was first separated into codons, and synonymous codons were randomly exchanged internally for each coding sequence for 6 consecutive times. Five independently shuffled replicas were generated for coding sequences. Shuffling was carried out separately for each coding sequence.

Assessment of TBP:
To assess TBP we determined frequency distribution of distances (FDD) between triplets as earlier described **[9]**. In brief, all positions of a given triplet were determined and distances (measured in bases) between successive triplets were calculated. The number of occurrences (frequency) at each distance was then counted.

Clustering of same-phase triplets:
To estimate the degree of clustering of same-phase triplets, we determined the number of triplets that were engaged in clustering in a head-to-tail collage of all analyzed coding sequences using non-overlapping windows of ten thousand bases each. Additionally, we determined clustering in the three reading frames identifying the size of each cluster (i.e. number of triplets per cluster); those determinations did not require the use of windows because all triplet clusters were quantified independently **[8]**.

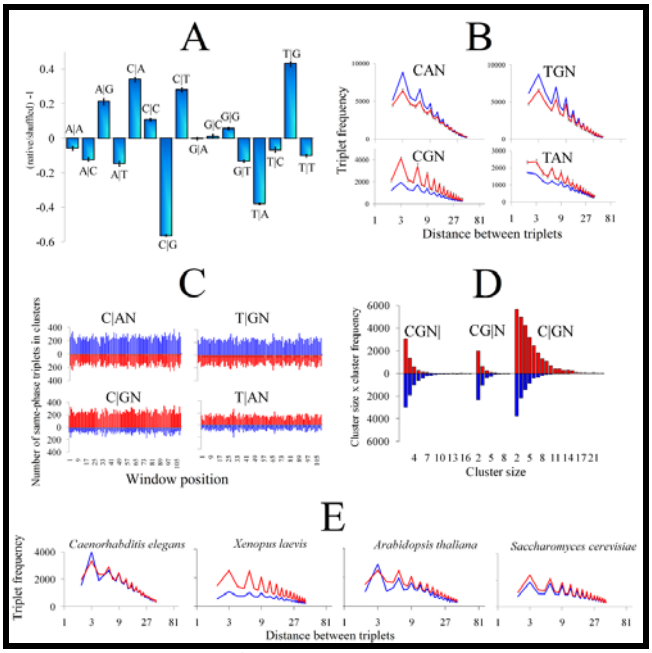

Figure 1: A) Assessment of intercodon dinucleotide tendencies. Tendencies were estimated by comparing native dinucleotide frequencies at codon junctions (e.g. A|G) against those in synonymous-codon-shuffled sequences. The y-axis shows ratios of native to synonymous-codon-shuffled frequencies minus the unit (-1), i.e. (native/shuffled) -1. Therefore positive values indicate intercodon preference while negative values indicate intercodon avoidance. Error bars represent 3 standard deviations (n=5). **B)** Analysis of 3-base periodicity by FDD of triplet formulas CAN, TGN, CGN and TAN in native and synonymous-codon-shuffled sequences. The x-axis is presented in logarithmic form to aid visualization of differences. Note that FDD does not discriminate between the different triplet phases. The blue line is for the native sequence and the red line is for the synonymous-codon-shuffled sequence. In the upper panels TBP is more intense in the native than in the shuffled sequence while in the lower panels dominance is inverted and TBP is more intense for the synonymous-codon-shuffled sequence than for the native one. In all cases error bars (on red line) represent 3 standard deviations. **C)** Clustering of same-phase triplets. The upper two panels correspond to triplet formulas CAN and TGN that displayed more intense 3-base periodicity in native sequences than in shuffled controls **(Figure 1B, upper panels)** with total frequencies of 25,521 and 15,454 ± 121 for C|AN and 27,047 and 17,906 ± 69 for T|GN and the lower two panels correspond to triplet formulas CGN and TAN that displayed less intense 3-base periodicity in native sequences than in shuffled controls **(Figure 1B, lower panels)** with total frequencies of 9,385 and 26,985 ± 71 for C|GN and 8,668 and 16,361 ± 63 for T|AN. Note that values above and below zero in the x-axis are both positive; hence in all cases the abundance and length of vertical lines are proportional to the number of same-phase triplets engaged in clusters as indicated in the y-axis. In the bottom of graphs the position of non-overlapping 10,000-bp windows is shown. **D)** Determination of same-phase triplet clustering in the three reading frames. The case for CGN is presented. In the x-axis the size of each cluster (number of triplets in each cluster) is shown. Clustering is presented for CGN|, CG|N and C|GN as indicated above each column set. Note that in the y-axis values above and below the x-axis are positive so that in both cases the column length is proportional to the number of triplets in clusters, i.e. cluster size x cluster frequency. Red columns above the x-axis are for the synonymous-codon-shuffled sequence and blue columns below the x-axis are for the native sequence. **E)** Analysis of 3-base periodicity of triplet CGN in coding DNA of the indicated organisms. With a blue line TBP patterns for native sequences are shown while patterns for synonymous-codon-shuffled sequences are shown with a red line. As in other cases the x-axis is presented in logarithmic form to help visualization and frequency values are shown in the y-axis.

Discussion:
Intercodon dinucleotide frequencies in native and randomized sequences:
We found higher intercodon T|G, C|A A|G and C|T (wherein the vertical line indicates the boundary between codons) frequencies in native sequences as compared to controls. Conversely, we found lower frequency of C|G and T|A in native sequences than in randomized controls. These results are summarized in **Figure 1A**, wherein bars represent the ratios of native dinucleotide frequency to synonymous-codon-shuffled dinucleotide frequency minus one (-1). The tendencies shown in **Figure 1A** should not be taken as equal to the long-known global dinucleotide tendencies in the genome **[17, 18]**, even though those tendencies partially agreed with ours, especially in the marked reduction of C|G; however, other dinucleotide tendencies did not coincide and could even go in opposite directions. Irrespective of it, coincidences are suggestive of intercodon tendencies being generated by the same mechanisms that affect e.g. CpG dinucleotides in the whole genome **[19]**.

Analysis of TBP and correlation with intercodon dinucleotide frequencies and with clustering:
According to results shown in **Figure 1A**, high or low intercodon dinucleotide frequencies could affect TBP differently. We therefore, investigated them separately by applying frequency distribution of distance (FDD). To study higher than random intercodon dinucleotide frequencies, we elected CA- and TG-related triplets, while for lower than random intercodon dinucleotide frequencies, we chose CG- and TA-related triplets. To reduce the number of calculations, we computed FDD for the triplet formulas CAN and TGN and CGN and TAN, wherein N jointly represents A, C, G and T. We found a correlation between the intensity of TBP and the type of analyzed triplet formula, so that TBP for CAN **(Figure 1B, panel A)** and TGN **(Figure 1B, panel B)** was more intense in native than in synonymous-codon-shuffled sequences and this coincided with preference for the intercodon dinucleotides C|A and T|G **(Figure 1A)**. Conversely, TBP for CGN **(Figure 1B, panel C)** and for TAN **(Figure 1B, panel D)** was less intense in native than in the shuffled sequence, which concurred with lower frequencies of C|G and T|A **(Figure 1A)**.

We subsequently investigated clustering of C|AN, T|GN, C|GN and T|AN. Results in **Figure 1C** demonstrate more frequent clustering of C|AN and T|GN triplets in native sequences than in their codon-shuffled counterparts. In contrast, clustering of same-phase triplets C|GN and T|AN was less frequent in native sequences than in shuffled controls. Therefore, in the native coding sequence TBP was more intense when the related intercodon dinucleotide was preferred and less intense, when it was avoided. This was just as predicted by the model for TBP **[8]**. Because of the formal possibility that other reading frames could contribute to the proposed effect, we quantified triplet clustering also in the two other possible reading frames, i.e. CG|N and CGN|. It should be noticed, however, that given the conservation of position 1 and 2 in codons, these calculations might reveal only potential clustering of codons. As expected, contrasts in clustering of CG|N and CGN| between native and

shuffled sequences were much weaker (data not shown). Quantification of the variously sized same-phase triplet clusters in the three possible CGN reading frames confirmed those findings **(Figure 1D)**, which demonstrate that clustering in the C|GN reading frame was more frequent in the randomized sequence than in the native one. That is, the two other reading frames did account for the effect over TBP.

To determine how different intercodon dinucleotide tendencies in other eukaryotes affected TBP, we contrasted native TBP patterns against those in synonymous-codon-shuffled coding sequences of *Caenorhabditis elegans*, *Xenopus laevis*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. The observed differences in TBP for the triplet CGN **(Figure 1E)** fully agreed with pre-existing C|G tendencies. Hence in *C. elegans* there was only marginal preference for C|G and this produced almost no difference in TBP for the triplet CGN **(Figure 1E)**. In contrast, C|G was strongly avoided in *X. laevis* and as a consequence CGN triplet TBP was more intense in the shuffled (**Figure 1E,** red line) than in the native *X. laevis* coding sequence (**Figure 1E,** blue line); this tendency was very similar to that observed in human coding DNA **(Figure 1B)**. Finally, in the native coding DNA of *S. cerevisiae* and in *A. thaliana* there was also C|G avoidance and this resulted in less intense TBP in the native sequence than in the shuffled one **(Figure 1E)**. We also briefly explored the effect of intercodon dinucleotides in 63 type I dengue virus coding sequences, in 244 HIV env protein coding sequences and in just over one thousand *Escherichia coli* ORFs. Fully compatible results were obtained for these organisms with similarity in CGN TBP and C|G clustering between the tested viruses and human (data not shown).

Potential biological relevance:

TBP may have a physiological consequence in the cell, which according to our results could involve intercodon dinucleotides, because it has recently been shown that 3-base periodicity and codon usage in yeast may be correlated with gene expression at the level of transcription elongation **[15]**. If transcription elongation was affected by TBP its variability among different eukaryotes would suggest different transcription elongation needs. The C|G intercodon dinucleotide may additionally be needed for gene expression. Therefore, removal of all CpG dinucleotides in the green fluorescent protein (GFP) led to a significant reduction in its expression via transcriptional attenuation **[20]**. We calculated that 80% of all removed CpG in GFP were C|G intercodon dinucleotides. Hence a low frequency of intercodon C|G and perhaps by extension a low level of C|GN clustering may be required for gene expression. Moreover, the similarity in C|GN clustering between dengue and HIV viruses and human coding DNA could suggest that viral genomes have adapted their transcriptional needs according to those of their host.

Conclusion:

Given the connection between TBP, intercodon dinucleotides and gene expression it is possible that details in TBP patterns will help reveal gene properties in a simple and expedite way, but techniques finer than the ones here employed need to be applied/developed for that purpose.

References:

[1] Li W. *Computers Chem.* 1997 **21**: 257 [PMID: 9415988]
[2] Gutiérrez G *et al. J Theor Biol.* 1994 **167**: 413 [PMID: 8207954]
[3] Eskesen ST *et al. BMC Mol Biol.* 2004 **5**: 12 [PMID: 15315715]
[4] Mena-Chalco JP *et al. IEEE/ACM Trans Comput Biol Bioinform.* 2008 **5**: 198 [PMID: 18451429]
[5] Yin C & Yau ST. *J Theor Biol.* 2007 **247**: 687 [PMID: 17509616]
[6] Wang L & Stein LD. *BMC Bioinformatics.* 2010 **11**: 550 [PMID: 21059240]
[7] Ma L *et al. BMC Genomics.* 2010 **11**: 416 [PMID: 20602772]
[8] Sánchez J & López-Villaseñor I. *FEBS Lett.* 2006 **580**: 6413 [PMID: 17097640]
[9] López-Villaseñor I *et al. Biochem Biophys Res Commun.* 2004 **325**: 467 [PMID: 15530416]
[10] Fedorov A *et al. Nucleic Acids Res.* 2002 **30**: 1192 [PMID: 11861911]
[11] Moura G *et al. PLoS ONE.* 2007 **2**: e847 [PMID: 17786218]
[12] Duan J & Antezana MA. *J Mol Evol.* 2003 **57**: 694 [PMID: 14745538]
[13] Stoletzki N & Eyre-Walker A. *Mol Biol Evol.* 2007 **24**: 374 [PMID: 17101719]
[14] Saunders R & Deane CM. *Nucleic Acids Res.* 2010 **38**: 6719 [PMID: 20530529]
[15] Trotta E. *PLoS ONE.* 2011 **6**: e21590 [PMID: 21738721]
[16] Pride DT *et al. Genome Res.* 2003 **13**: 145 [PMID: 12566393]
[17] Nussinov R. *J Mol Biol.* 1981 **149**: 125 [PMID: 6273582]
[18] Nussinov R. *J Biol Chem.* 1981 **256**: 8458 [PMID: 6943145]
[19] Pfeifer GP. *Curr Top Microbiol Immunol.* 2006 **301**: 259 [PMID: 16570852]
[20] Bauer AP *et al. Nucleic Acids Res.* 2010 **38**: 3891 [PMID: 20203083]

Edited by P Kangueane

Citation: Sánchez. Bioinformation 6(9): 327-329 (2011)