1192–1197 *Nucleic Acids Research, 2002, Vol. 30, No. 5*

Regularities of context-dependent codon bias in eukaryotic genes

Alexei Fedorov*, Serge Saxonov and Walter Gilbert

Department of Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received October 18, 2001; Accepted January 8, 2002

ABSTRACT

Nucleotides surrounding a codon influence the choice of this particular codon from among the group of possible synonymous codons. The strongest influence on codon usage arises from the nucleotide immediately following the codon and is known as the N_1 context. We studied the relative abundance of codons with N_1 contexts in genes from four eukaryotes for which the entire genomes have been sequenced: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana*. For all the studied organisms it was found that 90% of the codons have a statistically significant N_1 context-dependent codon bias. The relative abundance of each codon with an N_1 context was compared with the relative abundance of the same 4mer oligonucleotide in the whole genome. This comparison showed that in about half of all cases the context-dependent codon bias could not be explained by the sequence composition of the genome. Ranking statistics were applied to compare context-dependent codon biases for codons from different synonymous groups. We found regularities in N_1 context-dependent codon bias with respect to the codon nucleotide composition. Codons with the same nucleotides in the second and third positions and the same N_1 context have a statistically significant correlation of their relative abundances.

INTRODUCTION

The nucleotide composition of a gene coding sequence (CDS) is non-random. First, CDS non-randomness is engendered by the information needed to code for the protein primary structure. Second, CDS non-randomness is effected by the preferences in the choice of synonymous codons representing the same amino acid, the so-called codon bias. In addition to codon bias, neighboring nucleotides surrounding a codon influence the choice of this codon from the synonymous group. This phenomenon is known as context-dependent codon bias (CDCB) (1–3). The most important nucleotide determining CDCB is the first one following the codon (4,5) and is known as the N_1 context. Context N_i stands for the next *i* nucleotide after the codon, according to the notation of Berg and Silva (4). CDCB has been examined to a much lesser extent than codon bias itself. Regularities in CDCB have been described only fragmentally for a small portion of codons (4,6,7) while the whole picture of CDCB is unknown. Some cases of CDCB could be explained by the bias in the sequence composition of the entire genome. For example, human codons with a C nucleotide in the third position and with a G nucleotide as the N_1 context are significantly under-represented. The major reason for this under-representation is a 4-fold deficiency of the CG dinucleotides characteristic of the entire human genome, due to the methylation of cytosines within CG sites.

Information on codon bias and CDCB is very important for the improvement of gene-finding algorithms. All the available programs for gene prediction are far from perfect. In fact, we still do not have a good estimate of the number of genes in the human genome, as was shown in a recent report by Hogenesch *et al*. (8). A major part of gene prediction programs is based on the computation of characteristic non-randomness of coding and non-coding sequences. In this computational differentiation of genomic sequences on coding and non-coding pieces, the contribution of codon bias and CDCB is essential. Besides practical significance, knowledge of the CDCB is important in understanding the biological fundamentals of codon bias.

In this paper we have pursued several aims: (i) to carry out a thorough statistical analysis of the distribution of all codons with N_1 context in complete gene sets of different evolutionarily divergent species; (ii) to examine the extent of the influence of genomic sequence non-randomness on CDCB; and (iii) to expose possible regularities in CDCB. We analyzed CDCB by computing *R* values, where the *R* value represents the relative abundance for a codon $\underline{uvw}$ with N_1 context *n* computed as the ratio $R(\underline{uvw}{\sim}n) = F(\underline{uvw}{\sim}n)/[F(\underline{uvw})F(n)]$. $F(\underline{uvw})$ denotes the frequency of the codon $\underline{uvw}$ (*u*, *v*, *w* and *n* are the nucleotides a, g, t and c, and the codon is underlined), $F(n)$ is the frequency of nucleotide *n* in the N_1 context and $F(\underline{uvw}{\sim}n)$ is the frequency of the codon with the *n* context. Here and elsewhere the tilde character (~) separates codons (underlined) or oligonucleotides (non-underlined) from their mononucleotide context. The local non-randomness of the genome nucleotide composition was measured by the same approach by computing *r* values, the relative abundance of tri- di- and mononucleotide *y* with mononucleotide context *n*. They were calculated as the ratio $r(y{\sim}n) = F(yn)/[F(y)F(n)]$, where $F(y)$ denotes the

*To whom correspondence should be addressed. Tel: +1 617 495 0560; Fax: +1 617 496 4313; Email: afedorov@fas.harvard.edu
Present address:
Serge Saxonov, Stanford Medical Informatics, 251 Campus Drive, Medical School Office Building X-215, Stanford, CA 94305, USA

frequency of the oligonucleotide *y*, $F(n)$ the frequency of nucleotide *n* and $F(yn)$ the frequency of oligonucleotide *yn*. By comparing $R(\underline{uvw}{\sim}n)$ values computed for coding sequences with $r(uvw{\sim}n)$, $r(uw{\sim}n)$ and $r(w{\sim}n)$ values for the genomic sequences, we found that in some cases CDCB could be a consequence of the nucleotide composition characterized for the entire genome (genome bias). Nonetheless, in ~35–55% of the cases CDCB could not be explained by the genomic bias. In addition, we found regularities in the CDCB with respect to the nucleotide composition of codons and N_1 context. Our data support the hypothesis that the primary reason for codon bias and CDCB is selection for the accuracy of protein synthesis.

MATERIALS AND METHODS

Nucleotide samples

Genomic and CDS sequences of *Drosophila melanogaster*, *Caenorhabditis elegans* and *Arabidopsis thaliana* were downloaded from GenBank release 119 (9) (ftp: ncbi.nlm.nih.gov) as *.fna and *.ffn files, respectively. As a result, the entire genomic and CDS sequences for *D.melanogaster* and *C.elegans* and all the sequences from chromosomes I, II and IV of *A.thaliana* were obtained. Coding sequences with internal stop codons and those that did not start with an ATG codon or end in stop codons were removed from the samples. In the end, we obtained 14 335 coding sequences (21.6×10^6 nt) from *D.melanogaster*, 14 502 coding sequences (18.1×10^6 nt) from *C.elegans* and 10 145 (13.8×10^6 nt) from *A.thaliana*.

Since *.fna and *.ffn files are not available for human sequences, we used human genome contigs (0.5×10^9 nt) obtained from the NCBI (www.nlm.nih.gov/Genomes/index.html) as a source of genomic sequences. The human CDS sample was as published previously (intron-containing plus intronless samples) representing 782 genes (1.1×10^6 nt) (10).

The independent intron-containing CDS sample of *D.melanogaster* (IC-CDS) was obtained from our Exon-Intron Database (11). We removed all coding sequences with multiple duplications and purged the sample at the 20% amino acid similarity level as described in Fedorov *et al.* (10). Finally, the IC-CDS sample of *D.melanogaster* contained 505 coding sequences (0.67×10^6 nt). The entire set of *D.melanogaster* genes was randomly subdivided into two equally sized samples to generate random subsets 1 and 2.

Calculations of relative abundance of codons with context

For each codon x_i coding for amino acid *X* we computed $M_X(x_i{\sim}n)$, the number of occurrences of codon x_i with nucleotide *n* in the N_1 context in the CDS samples. To eliminate the bias in non-random associations of neighboring amino acids in a protein, we analyzed each group of synonymous codons separately. Based on the $M_X(x_i{\sim}n)$ table obtained we calculated $M_X(x_i)$, the number of occurrences of codon x_i in the sample [$M_x(x_i) = \sum_{n=a,g,t,c} M_x(x_i{\sim}n)$], $M_X(n)$, the number of occurrences of nucleotide *n* following the codons representing amino acid *X* [$M_x(n) = \sum_{x_i \in X} M_x(x_i{\sim}n)$], and M_X, the total number of codons representing amino acid *X* [$M_x = \sum_{n=a,g,t,c} \sum_{x_i \in X} M_x(x_i{\sim}n)$].

We then calculated the relative frequency of codon x_i with context *n* within synonymous group *X* [$F_X(x_i{\sim}n) = M_X(x_i{\sim}n)/M_X$], the relative frequency of codon x_i within group *X* [$F_X(x_i) = M_X(x_i)/M_X$] and the relative frequency of the context (*n*) within group *X* [$F_X(n) = M_X(n)/M_X$]. Finally, the relative abundance of codon x_i with context *n* [$R(x_i{\sim}n)$] was calculated by the formula:

$$R(x_i{\sim}n) = F_X(x_i{\sim}n)/[F_X(x_i)F_X(n)] \quad \mathbf{1}$$

Alternatively, we calculated the relative abundance of codon x_i with context *n* [$R(x_i{\sim}n)$] for the united pool of all codons with context by the formula:

$$R(x_i{\sim}n) = F_X(x_i{\sim}\mathrm{n})/[F_X(x_i)F(n)]$$

where $F(n)$ is now the frequency of nucleotide *n* in the first position of all codons. The results of this calculation are close to the results obtained using equation **1** and are present on our web page.

Standard deviation for $R(x_i{\sim}n)$

To estimate the significance of *R* values, we calculated standard deviations for each $R(x_i{\sim}n)$, using Monte Carlo simulations. For this purpose 100 independent random tests were applied. In each test, using a random number generator, we created $M_X^{rand}(x_i{\sim}n)$ distributions with the total number of codons within the synonymous group for each random sample equal to the corresponding number in the real sample ($M_X^{rand} = M_X$). The random number generator simulated the appearance of codons x_i with *n* context with the same frequencies as in the real sample, $F_X(x_i{\sim}n)$. For each random sample we calculated $R_k^{rand}(x_i{\sim}n)$, $k = 1, \ldots, 100$, according to equation **1**. The standard deviation for $R(x_i{\sim}n)$ was calculated by the formula:

$$\sigma(x_i{\sim}n) = \{\sum_{k=1}^{100} [R_k^{rand}(x_i{\sim}n) - \overline{R(x_i{\sim}n)}]^2/99\}^{1/2}$$

Calculations of relative abundance of mono-, di- and trinucleotides with context in the genomes

For the studied samples of genomic sequences we calculated the frequencies of each nucleotide $F(u)$, dinucleotide $F(uv)$, trinucleotide $F(uvw)$ and quadranucleotide $F(uvwn)$, where *u*, *v*, *w* and *n* are each one of the four nucleotides a, c, g and t. Then we calculated the relative abundances (*r* value) of the mono-, di- and trinucleotides with a single nucleotide context: $r(w{\sim}n) = F(wn)/[F(w)F(n)]$, for mononucleotide *w* with context *n*; $r(vw{\sim}n) = F(vwn)/[F(vw)F(n)]$, for dinucleotide *vw* with context *n*; $r(uvw{\sim}n) = F(uvwn)/[F(uvw)F(n)]$, for trinucleotide *uvw* with context *n*.

The *r* value of a mononucleotide with context $r(w{\sim}n)$ represents the so-called genomic signature, introduced by Karlin and Burge. (12).

Ranking statistics

We divided groups of synonymous codons into three types, based on their size and nucleotide composition in the third variable position. Type I was composed of those groups that contained four codons: the Ala, Gly, Leu(c) = [cta, ctc, ctg, ctt], Pro, Arg(c) = [cga, cgc, cgg, cgt], Ser(t) = [tca, tcc, tcg, tct], Thr and Val groups. Type II was composed of the groups containing two codons with pyrimidines in the third variable position: the Cys, Asp, Phe, His, Asn, Ser(a) = [agc, agt] and Tyr groups. Type III was composed of the groups containing two codons with purines in the third variable position: the Glu, Lys, Leu(t) = [tta, ttg], Gln and Arg(a) = [aga, agg] groups. Each of the three 6-fold degenerate codon groups (Arg, Leu

and Ser) was divided into a group of four degenerate codons and a group of two degenerate codons based on the nucleotide in the first and second codon positions. For instance, we analyzed a 4-fold degenerate group of arginine codons [Arg(c): cga, cgc, cgg, cgt] with a C nucleotide in the first position and a 2-fold degenerate group of arginine codons [Arg(a): aga, agg] with an A nucleotide in the first position separately. CDCB was compared between groups of synonymous codons belonging to the same type using ranking statistics (13). For this purpose, within a group of synonymous codons representing amino acid X, each codon x_i with context n was given a rank on the basis of its relative abundance value $R(x_i{\sim}n)$. The codon with the context which had the minimal $R(x_i{\sim}n)$ value was given rank 1, the codon with the next smallest $R(x_i{\sim}n)$ value was given rank 2, and so on. Then, homologous pairs (A_i and B_i) from synonymous groups A and B, which have the same nucleotide at the third codon position and the same nucleotide in the N_1 context, were compared by counting the absolute difference of their ranks $d_i^{AB} = |\text{rank}(A_i) - \text{rank}(B_i)|$. Finally, to obtain D values all d_i^{AB} values were summed:

$$D^{AB} = \sum_{i=1}^{16} d_i^{AB} \qquad \textbf{2a}$$

for type I synonymous groups or

$$D^{AB} = \sum_{i=1}^{8} d_i^{AB} \qquad \textbf{2b}$$

for synonymous groups of types II and III.
We used this D^{AB} value to measure the difference of CDCB between groups A and B.

We also performed ranking statistics for ranks which were normalized by the genomic signature $r(w{\sim}n)$ value. In this case, we calculated the normalized relative abundance values as $R'(\underline{uvw}{\sim}n) = R(\underline{uvw}{\sim}n)/r(w{\sim}n)$, where u, v and w are nucleotides and n is the context of the codon $\underline{uvw}$. We then computed ranks on the basis of the R' values.

The distribution of the D values of ranking statistics for two groups with non-correlated elements was simulated on a computer for 100 000 groups of 16 or 8 random elements, to which ranks were randomly attributed (see Fig. 3).

All calculations were performed by Perl scripts on a Pentium III computer running LINUX. The entire set of our data is available from our web site: www.mcb.harvard.edu/gilbert/cdcb.

RESULTS

An example of the calculated R values, the relative abundance of codons with N_1 contexts, obtained for the entire set of *D.melanogaster* genes is shown in Figure 1. The complete list of R values for all studied CDS samples of four species are presented on our web page: www.mcb.harvard.edu/gilbert/cdcb. The data show that 90% of codons with N_1 context have a statistically significant bias, since their R values differ from 1 by more than 3 SD. Fifty-five percent of the codons $\underline{uvw}$ with N_1 context n from the entire set of *Drosophila* genes have the following properties: (i) the $R(\underline{uvw}{\sim}n)$ value differs by >3 SD from all of the $r(uvw{\sim}n)$, $r(vw{\sim}n)$ and $r(w{\sim}n)$ values representing the genomic bias (relative abundance in the entire genome of the trinucleotide uvw, dinucleotide vw and nucleotide w with the n context, respectively); and (ii) the CDCB is larger than or opposite to the genomic bias. Therefore, the described

codon-context (*uvw* ~*n*)	CDCB R(*uvw* ~*n*)	standard *deviation* σ(*uvw* ~*n*)	genome bias r(*uvw* ~*n*)	genome bias r(*vw* ~*n*)	genome bias r(*w* ~*n*)	rank (normalized by r(*w* ~*n*))
ALA-group						
gct ~*a*	0.553	*0.004*	0.578	0.618	0.756	3
gct ~*c*	1.320	*0.005*	1.030	1.050	0.907	16
gct ~*g*	1.168	*0.004*	1.577	1.317	1.127	9
gct ~*t*	0.967	*0.007*	0.976	1.109	1.216	6
gcc ~*a*	1.425	*0.002*	1.342	1.240	1.130	12
gcc ~*c*	0.779	*0.003*	0.948	1.014	1.050	4
gcc ~*g*	0.745	*0.002*	1.019	0.945	0.932	7
gcc ~*t*	1.132	*0.004*	0.683	0.789	0.888	13
gca ~*a*	0.777	*0.004*	1.072	1.117	1.218	1
gca ~*c*	1.094	*0.006*	0.907	0.976	0.855	14
gca ~*g*	1.085	*0.004*	1.244	1.040	0.889	11
gca ~*t*	1.079	*0.008*	0.816	0.873	0.972	10
gcg ~*a*	0.656	*0.004*	1.033	1.032	0.907	2
gcg ~*c*	1.108	*0.006*	1.067	1.215	1.277	8
gcg ~*g*	1.354	*0.004*	1.234	1.062	1.046	15
gcg ~*t*	0.649	*0.006*	0.746	0.762	0.853	5
PRO-group						
cct ~*a*	0.584	*0.012*	0.540	0.618	0.756	4
cct ~*c*	1.080	*0.013*	1.176	1.050	0.907	13
cct ~*g*	1.192	*0.012*	1.261	1.317	1.127	8
cct ~*t*	1.217	*0.016*	1.137	1.109	1.216	7
ccc ~*a*	1.599	*0.004*	1.247	1.240	1.130	16
ccc ~*c*	0.484	*0.003*	1.136	1.014	1.050	1
ccc ~*g*	0.863	*0.003*	0.856	0.945	0.932	6
ccc ~*t*	1.114	*0.005*	0.760	0.789	0.888	14
cca ~*a*	0.843	*0.004*	1.050	1.117	1.218	2
cca ~*c*	1.185	*0.004*	1.104	0.976	0.855	15
cca ~*g*	0.952	*0.004*	1.048	1.040	0.889	10
cca ~*t*	1.055	*0.007*	0.837	0.873	0.972	11
ccg ~*a*	0.641	*0.004*	0.974	1.032	0.907	3
ccg ~*c*	1.388	*0.004*	1.412	1.215	1.277	12
ccg ~*g*	1.111	*0.004*	0.992	1.062	1.046	9
ccg ~*t*	0.730	*0.005*	0.731	0.762	0.853	5
SER(t)-group						
tct ~*a*	0.702	*0.012*	0.669	0.618	0.756	6
tct ~*c*	1.002	*0.013*	1.051	1.050	0.907	11
tct ~*g*	1.266	*0.013*	1.249	1.317	1.127	12
tct ~*t*	1.012	*0.015*	1.111	1.109	1.216	4
tcc ~*a*	1.327	*0.003*	1.127	1.240	1.130	13
tcc ~*c*	0.666	*0.004*	0.968	1.014	1.050	1
tcc ~*g*	0.882	*0.003*	0.953	0.945	0.932	7
tcc ~*t*	1.140	*0.005*	0.930	0.789	0.888	16
tca ~*a*	0.978	*0.006*	1.163	1.117	1.218	2
tca ~*c*	1.084	*0.006*	0.875	0.976	0.855	15
tca ~*g*	0.856	*0.006*	0.982	1.040	0.889	8
tca ~*t*	1.172	*0.009*	0.941	0.873	0.972	14
tcg ~*a*	0.742	*0.004*	1.030	1.032	0.907	3
tcg ~*c*	1.360	*0.005*	1.218	1.215	1.277	10
tcg ~*g*	1.098	*0.003*	1.005	1.062	1.046	9
tcg ~*t*	0.747	*0.005*	0.806	0.762	0.853	5
THR-group						
act ~*a*	0.604	*0.012*	0.670	0.618	0.756	4
act ~*c*	1.156	*0.013*	0.981	1.050	0.907	16
act ~*g*	1.188	*0.012*	1.161	1.317	1.127	10
act ~*t*	1.137	*0.014*	1.227	1.109	1.216	7
acc ~*a*	1.401	*0.003*	1.240	1.240	1.130	14
acc ~*c*	0.687	*0.003*	1.044	1.014	1.050	1
acc ~*g*	0.788	*0.003*	0.930	0.945	0.932	6
acc ~*t*	1.119	*0.004*	0.778	0.789	0.888	15
aca ~*a*	0.985	*0.005*	1.180	1.117	1.218	5
aca ~*c*	1.041	*0.006*	1.009	0.976	0.855	13
aca ~*g*	0.917	*0.005*	0.881	1.040	0.889	8
aca ~*t*	1.121	*0.007*	0.901	0.873	0.972	11
acg ~*a*	0.679	*0.004*	1.096	1.032	0.907	2
acg ~*c*	1.330	*0.005*	1.188	1.215	1.277	9
acg ~*g*	1.252	*0.004*	1.015	1.062	1.046	12
acg ~*t*	0.645	*0.005*	0.753	0.762	0.853	3

Figure 1. Relative abundance of *D.melanogaster* codons with N_1 context and genomic oligonucleotides with context. Relative abundance of codons with N_1 context, R values were calculated using equation **1**. Relative abundance of genomic oligonucleotides with context, r values were calculated as described in Materials and Methods. The ranking system is also described in Materials and Methods.

cases of CDCB cannot be explained by non-random associations of neighboring nucleotides in the studied genome. For example, the alanine codon gct with a c context has a $R(\underline{gct}{\sim}c)$ value of 1.32 ± 0.005 (Fig. 1). The genomic signature of the tc dinucleotide shows that this dinucleotide is deficient in the *Drosophila* genome [$r(t{\sim}c) = 0.907$] and, therefore, cannot cause the excess of gct codons with a c context. The genomic

bias of the trinucleotide gct with a c context [r(gct~c) = 1.03] and the dinucleotide ct with a c context [r(ct~c) = 1.05] in the *Drosophila* genome cannot be the reason for the much larger excess of gct codons with a c context either. Similar results were observed for the other three species examined. Forty-two percent of *A.thaliana* codons, 37% of *C.elegans* codons and 31% of human codons have a statistically significant CDCB which cannot be explained by genomic bias. The low percentage observed for the human genome is due to the smallest size of the human CDS sample and, thus, the largest values for standard deviation.

Since we deal with sets of genes, the obtained samples of codons are likely to be inhomogeneous and depend on gene composition. Because of this, we compared the R values obtained for several independent sets of *Drosophila* genes. In addition to the entire sample of *Drosophila* genes described above, we examined: (i) an experimentally confirmed intron-contained non-redundant set of 505 *Drosophila* genes; and (ii) two random gene subsets (subsets 1 and 2) representing half the total number of *Drosophila* genes. The R values for each of the four *Drosophila* samples were very similar to each other (see our web site). In 95% of the cases $|R_i(\underline{uvw}\text{~}n) - R_j(\underline{uvw}\text{~}n)| < 3\sigma_{max}(\underline{uvw}\text{~}n)$, where $\underline{uvw}$ is a codon with context n, i and j represent different *Drosophila* CDS samples and σ_{max} is the maximal $\sigma_i(\underline{uvw}\text{~}n)$ and $\sigma_j(\underline{uvw}\text{~}n)$ standard deviation. In 99% of cases $|R_i(\underline{uvw}\text{~}n) - R_j(\underline{uvw}\text{~}n)| < 5\sigma_{max}(\underline{uvw}\text{~}n)$. Therefore, none of the results presented above are affected considerably by the gene sampling.

Ranking statistics

It is clear from Figure 1 that there are regularities in CDCB. If there is a strong bias for a particular codon $\underline{uvw}$ with nucleotide n in the N_1 context, then it is likely that a similar bias exists for other codons having the same nucleotide w in the third position and with the same n context. This correlation is usually strongest for codons with the same nucleotide v in the second position as well. For instance, the alanine codon gcc with a c context is deficient in *Drosophila* genes [R(gcc~c) = 0.779 ± 0.003; Fig. 1]. The corresponding Ser (tcc), Pro (ccc) and Thr (acc) codons with the same c context are also deficient [R(ccc~c) = 0.484 ± 0.003, R(tcc~c) = 0.666 ± 0.004 and R(acc~c) = 0.687 ± 0.003]. The described CDCB has not been caused by biological processes at the genomic level, since the cc dinucleotide is in excess in the *Drosophila* genome r(c~c) =1.05. Also the genomic bias for the corresponding di- and trinucleotides with a c context cannot explain the described CDCB (see Fig. 1).

It is sensible to start an investigation of CDCB regularities comparing codons from synonymous groups of the same sizes and similar nucleotide compositions. That is why we divided the synonymous groups into three types, all having the same sizes and nucleotide compositions at the third codon position (see Materials and Methods). We compared R values of codons belonging to synonymous groups of the same type only. The similarity in CDCB between groups of synonymous codons representing amino acids A and B was measured as the D^{AB} value calculated using equation **2a** or **2b**. The smaller the D^{AB} value, the stronger the similarity in CDCB between the A and B groups of synonymous codons. The calculated D values for *D.melanogaster* and *A.thaliana* (Fig. 2) are much smaller than the simulated D values for two groups of 16 or 8 random

A — *Drosophila melanogaster*

Type I codons with N_1-context (16 items)

	Lc	V	St	P	T	A	Rc	G
Lc	0	18	58	60	50	68	66	72
V	18	0	60	56	46	62	54	66
St	58	60	0	24	30	38	42	26
P	60	56	24	0	24	30	42	30
T	50	46	30	24	0	26	30	32
A	68	62	38	30	26	0	28	36
Rc	66	54	42	42	30	28	0	26
G	72	66	26	30	32	36	26	0

Type II codons with N_1-context (8 items)

	H	Y	N	D	F	C	Sa
H	0	4	6	4	20	6	2
Y	4	0	2	4	16	6	4
N	6	2	0	2	14	8	6
D	4	4	2	0	16	8	6
F	20	16	14	16	0	16	20
C	6	6	8	8	16	0	4
Sa	2	4	6	6	20	4	0

Type III codons with N_1-context (8 items)

	Lt	Q	K	E	Ra
Lt	0	8	14	12	20
Q	8	0	8	8	18
K	14	8	0	6	12
E	12	8	6	0	16
Ra	20	18	12	16	0

B — *Arabidopsis thaliana*

Type I codons with N_1-context (16 items)

	Lc	V	St	P	T	A	Rc	G
Lc	0	34	52	78	72	62	64	62
V	34	0	56	74	72	62	68	72
St	52	56	0	42	42	14	76	64
P	78	74	42	0	24	36	100	94
T	72	72	42	24	0	34	102	88
A	62	62	14	36	34	0	86	72
Rc	64	68	76	100	102	86	0	60
G	62	72	64	94	88	72	60	0

Type II codons with N_1-context (8 items)

	H	Y	N	D	F	C	Sa
H	0	10	10	6	14	8	10
Y	10	0	2	12	8	2	6
N	10	2	0	12	10	4	8
D	6	12	12	0	16	10	12
F	14	8	10	16	0	6	6
C	8	2	4	10	6	0	4
Sa	10	6	8	12	6	4	0

Type III codons with N_1-context (8 items)

	Lt	Q	K	E	Ra
Lt	0	18	10	18	16
Q	18	0	16	6	8
K	10	16	0	14	16
E	18	6	14	0	4
Ra	16	8	16	4	0

Figure 2. D^{AB} values measuring the CDCB difference between groups A and B of synonymous codons with N_1 context. Groups of synonymous codons are marked by the letters of the amino acids they code for. Groups are divided into type I, II and III, based on their size and nucleotide composition in the third variable position. The tables present D values of pairwise comparisons for all groups of synonymous codons belonging to the same type. Synonymous group pairs with similar nucleotide compositions (having the same nucleotide in the second codon position) are boxed. Most frequently the minimal D values are located inside boxes and correspond to groups with similar nucleotide compositions. D values calculated for (**A**) *D.melanogaster* genes and (**B**) *A.thaliana* genes.

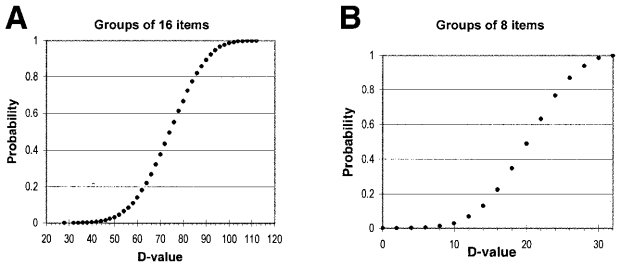

Figure 3. Cumulative probability distributions of D values for groups with randomly assigned ranks. The graphs can be used to assess the significance of D values. Specifically, the y-axis represents the probability that a pair of groups of 8 (**A**) or 16 (**B**) elements with randomly assigned ranks will have a D value less than or equal to the corresponding value on the x-axis.

elements shown in Figure 3A and B, respectively. Figure 3A shows that, on average, for two groups of 16 elements to which ranks were assigned randomly, the D value is 74. The D value is <56 for only 10% of the random group pairs, <42 for 1% of the random group pairs, <36 for 0.1% of the pairs and <28 for 0.01% of the pairs. At the same time, all D values calculated for *Drosophila* type I groups of synonymous codons with the N_1 context (Fig. 2) are less than the average value of 74 for the random groups. In seven of the 28 cases of pairwise comparison of *Drosophila* type I synonymous codon groups the D values are ≤28 (by chance, each occurrence has a probability of 10^{-4}). And in 20 of 28 cases these D values are <58 (by chance, each occurrence has a probability of 0.1). Similar ranking statistics

results were obtained for the type II and III groups of *D.melanogaster* synonymous codons, containing two codons per group and hence eight codons with an N_1 context. The data for ranking statistics between random groups of eight elements is shown in Figure 3B. The average *D* value for two random groups of eight elements is 20. A *D* value ≤12 was found for only 10% of the random group pairs and ≤8 for 1% of the random group pairs. In 15 of 21 cases of pairwise comparison of *Drosophila* type II synonymous codon groups and in four of 10 cases of *Drosophila* type III codon groups the calculated *D* values were ≤8 (by chance, each occurrence has a probability of 0.01). It is important to stress that most frequently the observed similarity is strongest between groups of synonymous codons with similar nucleotide compositions (those having the same nucleotide in the second position), which are boxed in Figure 2.

In summary, *Drosophila* codons with the same nucleotides in the second and third positions and the same N_1 context have a statistically significant correlation of their relative abundances. The same types of regularities of CDCB between codons with similar nucleotide compositions are detected for *Arabidopsis* genes (Fig. 2B) and for *H.sapiens* and *C.elegans*. The complete set of our results on the ranking statistics obtained for different types of rank normalization for different species is shown on our web site (www.mcb.harvard.edu/gilbert/cdcb).

DISCUSSION

Codon bias and CDCB are particular manifestations of coding sequence non-randomness, which is utilized in many different cellular processes. The best known use is that of codon bias in achieving efficiency and accuracy in protein synthesis (14–19). Also, in eukaryotic cells, CDS non-randomness is utilized in the splicing process. For many genes, the earliest steps of splicing of the pre-mRNA transcripts start with the binding of a group of SR proteins to the exonic sequences (for reviews see 20,21). The specificity of binding of SR proteins to coding sequences and the avoidance of intronic sequences is due to distinct motifs within the CDS. There is evidence that selection of such motifs within exons modulates the CDS non-randomness (10). After splicing, export of mRNAs to the cytoplasm occurs in complex with different proteins, some of which are SR proteins (22–24). In the cytoplasm, mRNAs exist in tight association with many proteins as a mRNP complex. The association of mRNAs with a variety of proteins, from their appearance during transcription until the time of translation, requires many motifs within the coding sequences and, therefore, creates CDS non-randomness. These motifs regulate mRNA fate in the cell. Because these motifs are involved in many cellular processes, the whole picture of codon bias and CDCB is very diverse and sometimes inconsistent. That is why there are many non-congruent facts about codon bias and CDCB in the literature. For example: (i) eukaryotes have a negative correlation of codon bias with gene length, while prokaryotes have a positive correlation (16,25,26); (ii) the evidence that over-represented codon pairs are translated slower than under-represented pairs (27) contradicts the theory that codon bias and CDCB increase the speed of translation (17–19,28); (iii) the observed congruency of in-frame and out-of-frame trinucleotide preference in *Drosophila* (6) is contrary to the idea that CDS periodicity is involved in the translation frame monitoring mechanism (29).

Here we would like to summarize the fundamental properties of codon bias and CDCB. First, codon bias is ubiquitous in all organisms, but evolutionarily remote species from different taxa have different patterns of codon bias (see Codon Usage Database, www.kazusa.or.jp/codon). Second, there are distinct regularities in the pattern of codon bias as described by the nucleotide composition of a codon via simple rules, shown by Karlin and Mrazek (5). We show that CDCB is ubiquitous for all organisms and that CDCB has distinct regularities. The CDCB regularities are also dependent on the nucleotide composition of a codon and, therefore, are in accordance with the codon bias regularities.

It is of interest to ascertain the underlying process that causes the appearance of codon bias and CDCB. There is an old hypothesis that the main reason for codon bias is translational efficiency, and the change in the relative concentrations of different isoaccepting tRNAs determines the codon bias. This hypothesis is based on the observation that there is a strong correlation between codon usage and the abundance of the corresponding tRNAs within each pool of isoaccepting tRNAs (18,19). We think that this scenario is highly unlikely, since it cannot explain: (i) the regularities of codon bias shown by Karlin and Mrazek (5); (ii) the existence of CDCB; and (iii) the regularities in CDCB shown in this paper. It is much more probable that the abundance of tRNAs is a consequence of and not a reason for codon bias. At the same time, variations of tRNA abundance should follow and stabilize the codon bias. Our results support the theory that the accuracy of protein synthesis on the ribosome is the primary reason for codon bias (14–16). Spatial interaction of ribosomal proteins with codon–anticodon RNA pairs inside the A and P sites of the ribosome could be preferable for particular codons with context. If such preferences exist they could be the primary reason for the regularities in codon bias and CDCB with respect to codon nucleotide composition.

ACKNOWLEDGEMENT

We thank Iraj Daizadeh for helpful discussions of the paper.

REFERENCES

1. Yarus,M. and Folley,L.S. (1984) Sence codons are found in specific contexts. *J. Mol. Biol.*, **182**, 529–540.
2. Shpaer,E.G. (1986) Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. *J. Mol. Biol.*, **188**, 555–564.
3. Gouy,M. (1987) Codon contexts in Enterobacterial and Coliphage genes. *Mol. Biol. Evol.*, **4**, 426–444.
4. Berg,O.G. and Silva,P.J.N. (1997) Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. *Nucleic Acids Res.*, **25**, 1397–1404.
5. Karlin,S. and Mrazek,J. (1996) What drives codon choices in human genes? *J. Mol. Biol.*, **262**, 459–472.
6. Antezana,M.A. and Kreitman,M. (1999) The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J. Mol. Evol.*, **49**, 36–43.
7. McVean,G.A.T. and Hurst,G.D.D. (2000) Evolutionary liability of context-dependent codon bias in bacteria. *J. Mol. Evol.*, **50**, 264–275.
8. Hogenesch,J.B., Ching,K.A., Batalov,S., Su,A.I., Walker,J.R., Zhou,Y., Kay,S.A., Schultz,P.G. and Cooke,M.P. (2001) A comparison of the

Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, **106**, 413–415.
9. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
10. Fedorov,A., Saxonov,S., Fedorova,L. and Daizadeh,I. (2001) Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Res.*, **29**, 1464–1469.
11. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
12. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.
13. Rice,A.J, (1995) *Mathematical Statistics and Data Analysis*, 2nd Edn. Duxbury Press, Belmont, CA.
14. Precup,J. and Parker,J. (1987) Missence misreading of asparagine codons as a function of codon identity and context. *J. Biol. Chem.*, **262**, 11351–11355.
15. Akashi,H. (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*, **136**, 927–935.
16. Eyre-Walker,A. (1996) Synonymous codon bias is related to gene length in *Esherichia coli*: selection for translational accuracy? *Mol. Biol. Evol.*, **13**, 864–872.
17. 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–r47.
18. Bennetzen,J.L. and Hall,B.D. (1982) Codon selection in yeast. *J. Biol. Chem.*, **257**, 3026–3031.
19. Ikemura,T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.*, **158**, 573–597.
20. Black,D.L. (1995) Finding splice sites within a wilderness of RNA. *RNA*, **1**, 763–771.
21. Berget,S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.*, **270**, 2411–2414.
22. Cole,C.N. (2000) mRNA export: the long and winding road. *Nature Cell Biol.*, **2**, E55–E58.
23. Siebel,C.W., Feng,L., Guthrie,C. and Fu,X.-D. (1999) Conservation in budding yeast of a kinase specific for SR splicing factors. *Proc. Natl Acad. Sci. USA*, **96**, 5440–5445.
24. Luking,A., Stahl,U. and Schmidt,U. (1998) The protein family of RNA helicases. *Crit. Rev. Biochem. Mol. Biol.*, **33**, 259–296.
25. 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.
26. Moriyama,E.N. and Powell,J.R. (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.*, **26**, 3188–3193.
27. Irwin,B., Heck,J.D. and Hatfield,G.W. (1995) Codon pair utilization biases influence translational elongation step times. *J. Biol. Chem.*, **270**, 22801–22806.
28. Robinson,M., Lilley,R., Little,S., Emtage,J.S., Yarranton,G., Stephens,P., Millican,A., Eaton,M. and Humphreys,G. (1984) Codon usage can affect efficiency of translation genes in *Escherichia coli*. *Nucleic Acids Res.*, **12**, 6663–6671.
29. Lagunez-Otero,J. and Trifonov,E.N. (1992) mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. *J. Biomol. Struct. Dyn.*, **10**, 455–464.