ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Biochemical and Biophysical Research Communications 304 (2003) 86–90

BBRC

www.elsevier.com/locate/ybbrc

Patterns of context-dependent codon biases

Anders Fuglsang*

Danish University of Pharmaceutical Science, Institute of Pharmacology, Universitetsparken 2, DK-2100 Copenhagen Ø, Denmark

Received 25 February 2003

Abstract

The association of codon context and codon usage was studied in seven bacteria as well as *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*. The association is strongest in magnitude closest to the codons of interest but there is apparently no rule about which of the two contexts is generally strongest associated to codon usage. In all bacterial species and in the intron-rich *Sch. pombe* it was furthermore observed from plots of χ^2 versus N that the wobble positions of codons in the proximity cause regular peaks both upstream and downstream. This observation is discussed in relation to a possible effect of mutational pressure on the association of codon usage and codon context. Absence of peaks corresponding to the wobble positions in the intron-poor *En. cuniculi*, and presence in *Sch. pombe*, may indicate that the role of introns in the context-dependent codon bias is negligible.

Keywords: Codon usage; Codon context; Mutational pressure; Contingency tables

In all species studied so far, codon usage has been shown to be far from random. Both selection and mutation are known to be causes of the non-random use of codons. As an example of the role for selection, the expressivity of genes is known to be a major factor in determination of codon usage phenomena, both eukaryotic as well as prokaryotic [1,2], as the codon usage has been adjusted in direction of the more abundant charged tRNAs [3,4]. To some extent, the composition of the DNA in an organism determines its codon usage, i.e., GC-rich synonymous codons are more frequently used in GC-rich organisms and vice versa [5–7]. However, it is also clear that the composition varies more locally as there are differences when comparing the leading strand of replication to the lagging strand, and differences in codon usage between genes of the two strands have been reported [8–10]. At the level of the gene, gene length is another factor that seems to be a determinant of composition and codon usage [2,11]. Even within genes can codon usage differences be observed. For example, it has been shown that the regional secondary structure of the gene product associates to codon usage in the genes [12,13]. In a few reports the 'codon context,' i.e., the composition around codons of interest, has been evaluated as a possible determinant of codon usage [14–19]. The N1 context (the nucleotide immediately downstream of codons of interest) is reported to be the strongest determinant of codon usage. Little is known about the cause of this but in a recent report it was concluded that selection seemed to play a major role in the context-dependent codon bias [19]. The purpose of this study was to evaluate the context-dependent codon bias and stretch the analysis to include many more context nucleotides than just the ones in the immediate neighbourhood, using another technique than that of the recent report by Federov et al. [19].

Materials and methods

Species in this study. The genomes of *Escherichia coli* (GenBank Accession No. NC000913), *Staphylococcus aureus* (NC002745), *Pseudomonas aeruginosa* (NC002516), *Bacillus subtilis* (NC000964), *Lactococcus lactis* (NC002662), *Streptococcus pyogenes* (NC002737), and *Mycobacterium tuberculosis* (NC00962) were downloaded and the protein encoding regions were extracted. Annotated sequences were included if they were considered valid, that is, had correct start and stop codons, an intact frame, and no internal stop codons or non-ACGT letters. In order to get to study how presence of introns may affect context-dependent codon usage it was decided also to include two eukaryotes in this study, *Encephalitozoon cuniculi* (chromosome

* Fax: +45-35-30-60-20.
E-mail address: anfu@dfh.dk.

0006-291X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0006-291X(03)00530-8

I–XI) and *Schizosaccharomyces pombe* (chromosome I–III). The former rarely has introns, whereas they occur frequently in genes of the latter. When genes with introns were parsed for analysis (see below) it was the joined sequence corresponding to the ribosomally processed mRNA that was parsed.

Contingency tables. We decided to study the context up to $N = 20$ nucleotides away from codons of interest, while at the same time avoiding gene boundary conflicts. Therefore codons were only included in this study if they were present at least 10 frames from the start codon and 10 frames from the stop codon. This condition is in principle not programmatically necessary but allows comparison of χ^2-values within species. For every amino acid having synonymous codons all codons were mapped along with the nucleotides from 1 to 20 positions upstream or downstream. Based on this, 40 contingency tables were constructed for each amino acid having synonymous codons.

The principle of contingency tabulation is comparison of observed counts with expected counts in a systematic manner. For a table with R rows and C columns, and where we have observed cell counts of $N_{\text{obs}}(c,r)$, the row total for the rth row is given by

$$T_{\text{row}}(r) = \sum_{c=1}^{C} N_{\text{obs}}(c,r)$$

and the column total of the cth column is:

$$T_{\text{col}}(c) = \sum_{r=1}^{R} N_{\text{obs}}(c,r).$$

The total table count T is thus

$$T = \sum_{c=1}^{C}\sum_{r=1}^{R} N_{\text{obs}}(c,r).$$

If we look at a the cell designated by (c,r) and calculate what the expected count is, then we know that in the row of this cell we have in total $T_{\text{row}}(n,c)$. Overall, we expect the fraction $T_{\text{col}}(c)/T$ to go into that particular column where the cell of interest is, so the expected count of cell (c,r) becomes:

$$N_{\text{exp}}(c,r) = \frac{T_{\text{row}}(r) \times T_{\text{col}}(c)}{T}.$$

The information about actual counts in the table and the expected counts in the table are then compared by an approximation, as the quantity

$$\sum_{c=1}^{C}\sum_{r=1}^{R} \frac{(N_{\text{obs}}(c,r) - N_{\text{exp}}(c,r))^2}{N_{\text{exp}}(c,r)}$$

asymptotically becomes χ^2-distributed (with $(C-1)\times(R-1)$ degrees of freedom) for $T \to \infty$. In practice, the conclusion drawn from a contingency table can be trusted when the table total is 40 or more [20]. In this study all contingency tables have totals of thousands because entire chromosomes are parsed, so the χ^2-inferred probabilities are likely to be extremely accurate, and the above quantity will in accordance with common practice be referred to as χ^2. A significance level of less than 5% was considered significant. A P-value of less than 5% indicates that the proportions of codons actually used differ between the four possible context letters.

All contingency tables for a particular amino acid and species encompass the information drawn from the exact same sample and therefore the χ^2-values of the contingency tables can be directly compared both upstream and downstream. For each amino acid, plots were constructed where χ^2-values were plotted against N, both upstream and downstream.

In addition hereto, the raw differences between observed counts and the expected counts for those cells of the contingency tables that had a codon ending in the same letter as the letter of the context were summed. This produces plots of difference sums as function of N. The magnitude of the difference sums (Δ) directly tells if there is an overrepresentation (Δ positive) or underrepresentation (Δ negative) of situations where the context letter is the same as the third letter of the codon of interest. The argument for doing so is given in the results section. The difference sums of all cells sum to exactly zero in contingency tables, in other words, the difference sum of a contingency table where we only include those cells where the context letter is *NOT* the same as the third letter in the codon of interest will handily sum up to exactly $-\Delta$.

Results

The amount of data generated with all amino acids and species in this study is so large that it is not possible to present them all here. We shall in the following present data for two bacteria *S. aureus* and *E. coli*, along with data for *En. cuniculi* and *Sch. pombe*. The two bacteria represent two extremes of the results observed with the bacteria mentioned in the Materials and Methods section. More specifically, the context-dependence for arginine codons was chosen for presentation, as they exemplify well the observations done with other amino acids having synonymous codons. The data for other species are available from the author upon request.

Fig. 1 shows a plot of χ^2-values as function of N for *E. coli* for arginine codons. The three horizontal lines on the figure depict the 5% level (lower), the 1% level, and the 0.1% level (upper). Three principally distinct aspects are visible from this figure. First, the association of context and codon usage is strongest in the closest proximity, simply because the highest χ^2-values are observed here. Next, the association is highly significant even up to a context distance of 20 nucleotides. Third, a 'sawtooth' pattern is observed for the χ^2-values, and the peaks occur exactly at positions corresponding to the wobble position (hereafter referred to as wobble peaks) of the codons of the context ($N = 1, 4, 7, \ldots$, upstream,

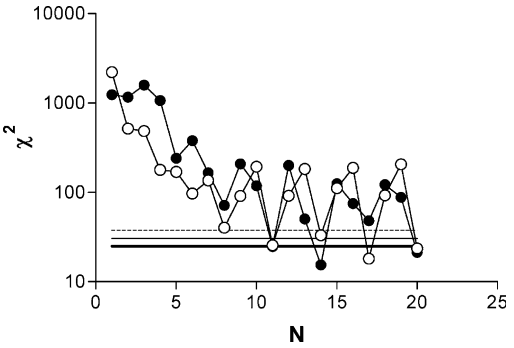

Fig. 1. The effect of codon context on codon usage, shown as contingency χ^2-values for nucleotides N positions upstream (○) or downstream (●) of arginine codons in *E. coli*. On the logarithmic axis, the three horizontal lines denote the probability levels of 5% (lower), 1 and 0.1% (upper). Note that the effects are stronger close to the codon and that the wobble positions of neighbouring codons give rise to peaking χ^2-values. The third letter of flanking codons thus has strong associations to codon usage.

$N = 3, 6, 9, \ldots$, downstream). In Fig. 2 similar data are shown for *S. aureus*. In this bacterium the wobble peaks are not visible, but the two other traits seen in *E. coli* are also visible here. The same plot for chromosome I–XI of *Sch. pombe*, a eukaryote with many introns (Fig. 3), is similar to that of *E. coli*, with very clear wobble peaks, whereas the wobble peaks are not very visible with the intron-poor *En. cuniculi* (Fig. 4), suggesting a minor role for introns in the context-dependent codon bias. If we look at the context-dependent usage of start and stop codons in *E. coli* (Fig. 5) then it can be seen that the stop codon seems to be more strongly associated to its context than the start codon is, judging from the χ^2-values.

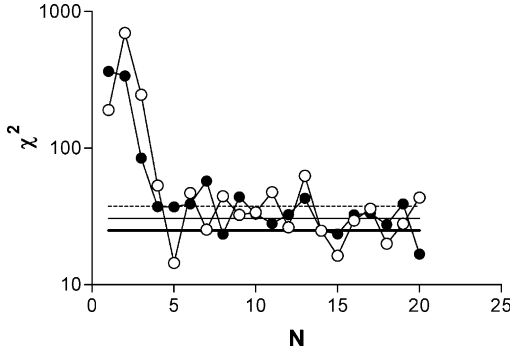

Fig. 2. Similar to Fig. 1, but shown for arginine codons in *S. aureus*.

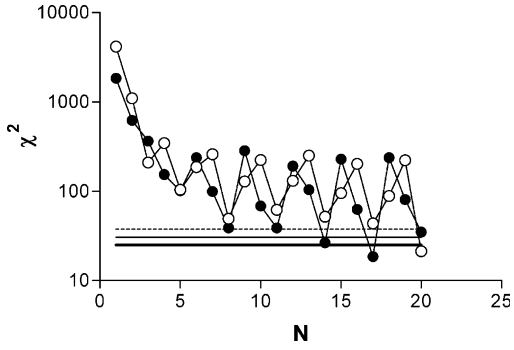

Fig. 3. Similar to Fig. 1, but shown for chromosome I–III in *Sch. pombe*. In this organism introns occur frequently.

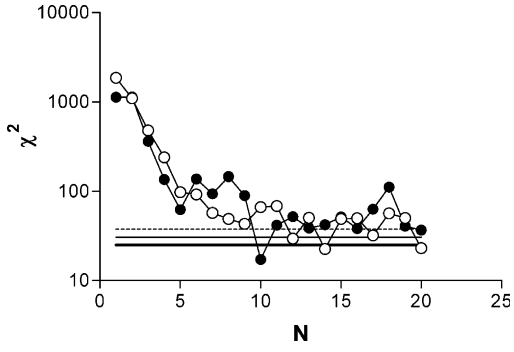

Fig. 4. Similar to Figs. 1–3, but shown for chromosome I–XI of *En. cuniculi*. This organism rarely has introns in its genes.

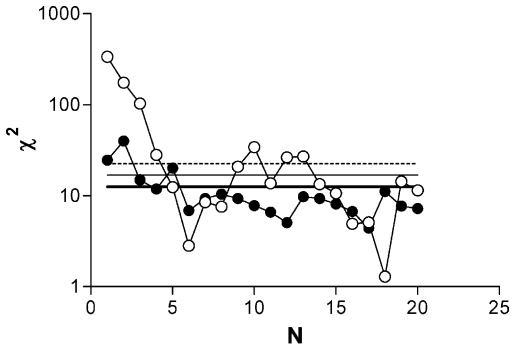

Fig. 5. The effect of codon context on start- and stop codon usage for nucleotides N positions upstream of stop codons (○) or downstream start of codons (●) in *E. coli*.

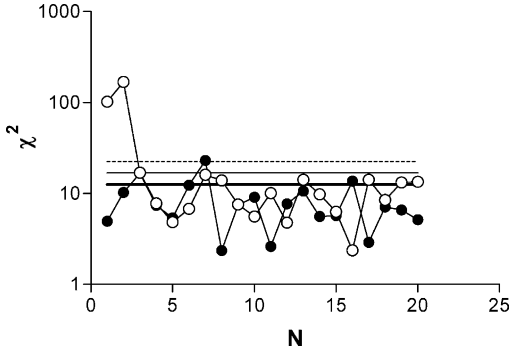

Fig. 6. The effect of context on stop codon usage for nucleotides N positions upstream of stop codons in *Sch. pombe* (●) and *En. cuniculi* (○).

The same holds true for other bacteria. In *Sch. pombe* the stop codons are not associated with high significance to its context, whereas it is in *En. cuniculi* for the last two nucleotides before stop codons (Fig. 6).

We hypothesize that the peaks corresponding to the wobble positions (Figs. 1 and 2, might originate from the directional mutational pressure. The wobble position is generally believed to be more prone to mutations because of the freedom here (for example, changing the third letter of the arginine codon CGA does not change the amino acid encoded by that codon). So if the directional mutational pressure were the cause of peaking χ^2-values, then we would expect that there would be an overrepresentation of codons having the same third letter as the letter in the context when the χ^2-peaks are observed, or, a directional mutational pressure will increase the likelihood of finding matching letters, especially in the wobble positions, because mutations here are often silent. Therefore we plotted the raw difference sums (Δ, sums of observed counts minus expected counts in the contingency tables), in cases where the context letter is the same as the third letter of the codon of interest, against N. An example of such a plot is given in Fig. 7, where Δ is plotted against N for arginine in *E. coli*. A similar sawtooth-like pattern is observed in

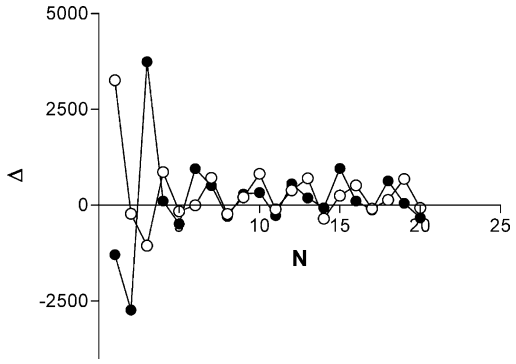

Fig. 7. *Escherichia coli* plots of difference sums (observed minus expected counts) from the contingency tables in cases where the third letter of the codon of interest is the same as the letter in the context where (○) indicate upstream contexts and (●) indicate downstream. Positive values of Δ occur when such cases are overrepresented. The figure indicates overrepresentation of such cases for the wobble positions in the context.

this figure, telling that there is an overrepresentation of context letters equal to the third letter of the arginine codons, when we look at the wobble positions in the context. In other words, the wobble peaks arise because the arginine codons tend to have the same third letter as the third letters of codons in the context. Exactly the same is observed for *Sch. pombe* (Fig. 8). This observation is thus compatible with the hypothesis that a

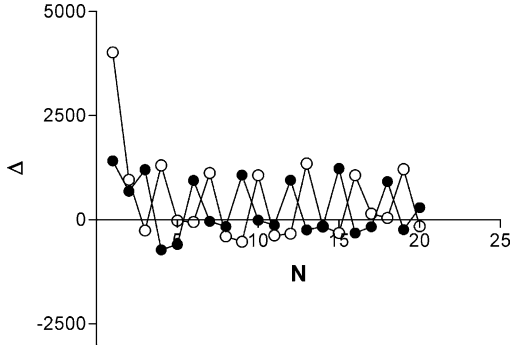

Fig. 8. Similar to Fig. 7, but for *Sch. pombe*.

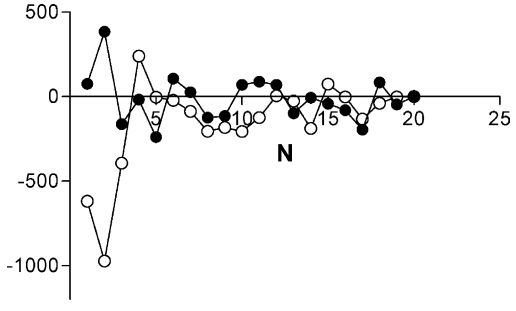

Fig. 9. Similar to Figs. 7 and 8, but for *S. aureus*. No pattern corresponding to the wobble positions is observed. See also Fig. 2.

directional mutational pressure plays an important role in the context-dependence for these species. However, Fig. 9 shows the same plot for *S. aureus*, and like in Fig. 2 no peaks corresponding to the context's wobble positions can be seen, suggesting a minor role for the mutational pressure in the context-dependent codon bias in this species. In *En. cuniculi* a weak tendency towards peaking Δ-values can be seen (not shown).

Discussion

The context-dependent codon bias has been studied and found to be strongest in magnitude closest to the codon, but with here there is little evidence in support of the N1 context being the stronger, as it was reported elsewhere [21]. The magnitudes of the wobble peaks (Figs. 1 and 3) seem smaller than those of the peaks occurring with the context closest to the codons of interest, indicating that no single force responsible for the context-dependent codon usage bias can explain the curves; an additional force seems to play role in the closest proximity which is stronger than the force acting farther away. A closer inspection of the χ^2-values causing the wobble peaks indicates that they in part are caused by overrepresentation of context nucleotides matching the third nucleotide of the codons on interest, which in our opinion strongly favours a role for the mutational pressure. But there is still the question why the χ^2-values are much greater—due to some other factor—in the closest proximity to codons. Several possibilities have been considered: it is well known that the codon bias is many organisms is under strong selectional influence, as it is adapted towards a maximization of translational efficiency [1,2]. A highly expressed gene is likely to display a set of codons corresponding to the more abundant synonymous tRNAs, thereby decreasing the risk of charged tRNAs becoming a limiting factor in the protein synthesis [3,4]. Looking at our figures it looks like the strongest effects of context-dependent codon bias in the closest proximity only extend 5–6 nucleotides away from codons of interest, equivalent to about two codons. The elongation process in ribosomes, both prokaryotic as well as eukaryotic, involves two places termed the aminoacyl-locus, where charged tRNAs arrive, and the peptidyl-locus, where the peptide chains are positioned. With (partial) imperfect Watson–Crick pairing being possible for tRNAs there is a risk of incorporation of the wrong amino acid. There is some experimental proof to suggest that codon bias in some species is adjusted towards minimization of this risk, so the observation of the context-dependent codon bias being strong up to 6 nucleotides around codons of interest might suggest that the codon bias arise from selection producing optimal combinations of codons. For example, if there is a selectional advantage of having imperfect Watson–Crick pairing at

exactly one of the two loci in the ribosome, then the codons might evolve towards certain combinations of codons compatible with the observations done here.

We did not observe any wobble peaks for the context at stop- or start codons (Figs. 5 and 6), but there were some effects of the context in the very close proximity to start- and stop codons. The lack of wobble peaks can be seen as a lack of association between the mutational pressure and choice of start- or stop codons.

The χ^2-values of contingency tables increase linearly with the table totals for a given set of proportions. So insignificant χ^2-values may, as with many other statistics, be caused by a low table total or because there is in fact no difference between the proportions in the table. Lack of wobble peaks could thus simply be observed because of too low table totals (the table total of these contingency tables is equal to the number of genes because each of them contains exactly one start- or stop codon). In yeast it has been shown that that there is some degree of start codon context effects on expression [22] and codon usage [16,17], and similar observations have been reported for other organisms such as amoeba [23] as well as *E. coli* [24–26]. If selection acts as claimed in this region to adjust the combinations of start codons and their downstream context towards optimal combinations for the desired (but not necessarily high) expression level, then that should in principle be reflected in the contingency tables, and this could account for the significant χ^2-values for the start context in Fig. 5.

The technique applied in the study of the phenomenon cannot be compared to the technique used elsewhere [19]. In fact, the two techniques are unlikely to measure the same. For example, in this study we get information about the context stretching up to 20 nucleotides both upstream and downstream with little opportunity to study the impact of selection on the context-dependent codon bias, while with the technique in [19], there is only retrieved information about the nucleotide combinations corresponding to the N1 context. A future objective could be to use both techniques on the same organisms and extend the technique of others to allow data retrieval for other contexts than just N1.

References

[1] M. Gouy, C. Gautier, Codon usage in bacteria: correlation with gene expressivity, Nucleic Acids Res. 10 (1982) 7055–7074.

[2] L. Duret, D. Muchiroud, Expression pattern and, surprisingly, gene length, shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*, Proc. Natl. Acad. Sci. USA 96 (1999) 4482–4487.

[3] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the relative occurence of the respective codons in protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, J. Mol. Biol. 21 (1981) 389–409.

[4] T. Ikemura, Codon usage and tRNA is unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.

[5] A. Pan, C. Dutta, J. Das, Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias, Gene 215 (1998) 405–413.

[6] G.E. Andersson, P.M. Sharp, Codon usage in the *Mycobacterium tuberculosis* complex, Microbiology 142 (1996) 915–925.

[7] H. Romero, A. Zavala, H. Musto, Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*, Gene 242 (2000) 307–311.

[8] M.J. McLean, K.H. Wolfe, K.M. Devine, Base compositional skews, replication origientation and gene orientation in 12 prokaryote genomes, J. Mol. Evol. 47 (1998) 691–696.

[9] P. Lopez, H. Phillipe, Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation, Life Sci. 324 (2001) 201–208.

[10] J.R. Lobry, N. Sueoka, Asymmetric directional mutation pressure in bacteria, Genome Biol. 3 (2002) 1–14.

[11] E.N. Moriyama, J.R. Powell, Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*, Nucleic Acids Res. 26 (1998) 3188–3193.

[12] M. Oresic, D. Shalloway, Specific correlations between relative synonymous codon usage and protein secondary structure, J. Mol. Biol. 281 (1998) 31–48.

[13] X. Tao, D. Dafu, The relationship between synonymous codon usage and protein structure, FEBS Lett. 434 (1998) 93–96.

[14] M. Bulmer, The effect of context on synonymous codon usage in genes with low codon usage bias, Nucleic Acids Res. 18 (1990) 2869–2873.

[15] G.A. McVean, G.D. Hurst, Evolutionary lability of context-dependent codon bias in bacteria, J. Mol. Evol. 50 (2000) 264–275.

[16] H. Miyasaka, The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*, Yeast 15 (1999) 633–637.

[17] H. Miyasaka, Translation initiation AUG context varies with codon usage bias and gene length in *Drosophila melanogaster*, J. Mol. Evol. 55 (2002) 52–64.

[18] B.R. Morton, B.G. So, Codon usage in plastid genes is correlated with context, position within the gene, and amino acid content, J. Mol. Evol. 50 (2000) 184–193.

[19] A. Fedorov, S. Saxonov, W. Gilbert, Regularities of context-dependent codon bias in eukaryotic genes, Nucleic Acids Res. 30 (2002) 1192–1197.

[20] W.G. Cochran, Some methods for strengthening the common χ^2-tests, Biometrics 10 (1954) 417–451.

[21] O.G. Berg, P.J. Silva, Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection, Nucleic Acids Res. 25 (1997) 1397–1404.

[22] A.C. Looman, M. Laude, U. Stahl, Influence of the codon following the initiation codon on the expression of the lacZ gene in *Saccharomyces cerevisiae*, Yeast 7 (1991) 157–165.

[23] J.D. Wuitschick, K.M. Karrer, Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in *Tetrahymena thermophila*, J. Eukaryot. Microbiol. 46 (1999) 239–247.

[24] C.M. Stenström, E. Holmgren, L.A. Isaksson, Cooperative effects by the initiation codon and its flanking regions on translation initiation, Gene 273 (2001) 259–265.

[25] J.P. Etchegaray, M. Inouye, Translational enhancement by an element downstream of the initiation codon in *Escherichia coli*, J. Biol. Chem. 274 (1999) 10079–10085.

[26] J.P. Etchegaray, M. Inouye, A sequence downstream of the initiation codon is essential for cold shock induction of cspB of *Escherichia coli*, J. Bacteriol. 181 (1999) 5852–5854.