Volume 285, number 2, 165–169 FEBS 09993 July 1991

ADONIS 0014579391006976

Minireview

Codon bias and gene expression

C.G. Kurland

Department of Molecular Biology, Uppsala University, Biomedical Center, Box 590, Uppsala S751 24, Sweden

Received 3 May 1991

The frequencies with which individual synonymous codons are used to code their cognate amino acids is quite variable from genome to genome and within genomes, from gene to gene. One particularly well documented codon bias is that associated with highly expressed genes in bacteria as well as in yeast; this is the so-called major codon bias. Here, it is suggested that the major codon bias is not an arrangement for regulating individual gene expression. Instead, the data suggest that this codon bias, which is correlated with a corresponding bias of tRNA abundance, is a global arrangement for optimizing the growth efficiency of cells. On the practical side, it is suggested that heterologous gene expression is not as sensitive to codon bias as previously thought, but that it is quite sensitive to other characteristics of the heterologous gene.

Codon bias; Aminoacyl-tRNA; Cell growth; mRNA stability; Protein stability

1. INTRODUCTION

The degeneracy of the genetic code enables the same amino acid sequences to be encoded and translated in many different ways. However, the alternative coding possibilities are not expressed in a purely random way. Rather, systematic bias of degenerate codon usage appears at different levels of genetic organization. These biases define what may be thought of as the phenotypes of genomes.

1. At the highest level are found characteristic ways for codon degeneracy to be exploited by different groups of organisms in what has been called 'genomic strategies' [1].

2. The codon sequences within a given genome may be locally biased: for example, discontinuities of base composition between local domains of animal genomes lead to a mosaic arrangement of codon preferences [2]. This means that the homologous sequences of the individual members of protein families may be coded in different ways within the same genome. A more subtle version of this is to be found in procaryotic genomes for which there is a gradient of codon preferences that is organized around the origin of replication [3]. Likewise, classes of genes within the same genome that are physiologically regulated to different expression levels may have class-specific codon preferences as for the so-called major codon preference [4].

Correspondence address: C.G. Kurland, Department of Molecular Biology, Uppsala University, Biomedical Center, Box 590, Uppsala S751 24, Sweden

3. There are intragenic codon biases. For example, in *Escherichia coli* as well as in *Sacharomyces cerevisiae* there is a codon bias in the initial sequences of genes which for major proteins is strikingly different from their downstream codon bias [5].

Here, I will focus on the major codon bias as well as on intragenic codon bias. One reason for this focus is that the greatest interest in codon bias has been aroused by its potential relevance to a practical problem. This concerns the efficiency with which a gene that has one codon bias can be expressed by a translation system that is adapted to a different codon bias. This problem of heterologous gene expression is one that unfortunately has been complicated by biases of yet another sort.

2. BIAS À LA MODE

As soon as a significant number of genes had been sequenced, it became accepted opinion that biased codon usage could regulate the expression levels of individual genes by modulating the rates of polypeptide elongation [4,6–10]. This opinion was reinforced by two lines of evidence. One consists of data suggesting that tRNA distributions of cells tend to follow the cognate codon frequencies of the mRNA pool [11–14]. The other is that the rates of polypeptide synthesis vary depending on the character of the codons being translated [14–17], as are the initial selection kinetics for tRNA ternary complexes [18]. Nevertheless, the intuition that there is a regulatory relationship between codon bias and the expression levels for individual genes that is mediated

by codon-specific rates of translation is doubtful in the extreme.

For example, a very poorly matched combination of codon biases consists of *Halobacterium halobium* with its genomic GC content close to 70% and *Escherichia coli* with its GC content close to 50%. The fact is that bacterio-opsin produced from the *H. halobium* gene is expressed at a high level in *E. coli*. Furthermore, a comparison of this expression level with that from a modified version of the same gene with more than half of its codons (142 out of 251) replaced by those preferred by *E. coli* reveals no significant differences [19]. Other examples of heterologous genes that are expressed at a high level in cells adapted to a different codon bias have been described [20]. The demonstrable point it that a heterologous gene is not necessarily expressed at a low level simply because it is made up of codons that are infrequently translated by the host cell.

There are other reasons why heterologous genes or synthetic genes may be poorly expressed. Two of these are evident in the experiments of Nassal et al. [19]. One is that the coding string corresponding to the N-terminal sequence of the protein is critical. If the beginning of the mRNA does not share a consensus sequence common to the genes of *E. coli*, its translation efficiency will be low for reasons discussed in the next section. Nassal et al. [19] solved this problem by appending their opsin sequence to an appropriate fragment containing the efficient consensus sequence for *E. coli*.

The other problem is that RNA or protein products expressed from heterologous or engineered genes may be unstable in the host cell. Indeed, Nassal et al. [19] observed that opsin is unstable in *E. coli*, but that the addition of a short polypeptide tail to the N-terminal sequence stabilized the polypeptide.

A similar cautionary tale concerns the stability of the mRNA produced by an engineered gene. One of the enduring myths of the field is that double stranded structures in mRNA will impede the progress of the translating ribosome and that accordingly, such structures can regulate protein expression levels [21,22]. Indeed, when sequences that contain putative double stranded structures were inserted into a *lacZ* sequence to test this notion, the protein expression level was reduced by more than a factor of ten [17]. However, controls revealed that the inserted sequences had destabilized the mRNA so that it was present at one-tenth its normal concentration in the bacteria. It seems that the double stranded structure provides an attractive site for a nuclease such as RNase 3. The moral of the story is that after the appropriate controls are done, what seemed to be an indication that double stranded inserts cause a reduction of translational efficiency simply vanishes.

There is also clear evidence that relatively small changes in mRNA sequences can have relatively large effects on mRNA stability [23,24]. Accordingly, any experiment that reveals a lowered protein expression level following replacement of a codon string is essentially uninterpretable unless measurements of mRNA levels and protein turnover are presented. The odd thing is that few of the authors describing experiments supporting the view that protein expression levels can be regulated by codon substitutions have bothered to measure the stabilities of the corresponding gene products.

In summary, there are experiments showing that proteins coded by seldomly translated codons are not necessarily expressed at low levels. Furthermore, there are no experiments showing that the rates with which codons are translated can regulate the expression level of the protein that they specify. Indeed, we would not expect such evidence to be forthcoming.

3. INTRAGENIC BIAS

If a cell were translating a single mRNA species, the rate at which the elongation process is carried out might influence the protein expression level if ribosomes are present in limiting amounts. This is so because the faster a polypeptide chain is completed, the more rapidly the ribosomes can return to initiate and complete another polypeptide chain. However, if a cell is processing hundreds of different mRNA species, changing the rate at which one of these is translated will not have a proportional effect on the expression level of the corresponding protein. This is so because after the ribosome completes the translation of one mRNA, we expect it to be sequestered most often by a different mRNA. Thus, in the absence of a mechanism to keep the ribosome on the same mRNA, the kinetic advantage resulting from rapid translation of one mRNA species is partitioned among all of the competing mRNA species. In other words, we expect the protein expression level to be determined only by the number of mRNA species expressed and the number of ribosomes that translate each mRNA.

In contrast, the rate of ribosome initiation of a mRNA might very well influence the expression level by determining the number of ribosomes that translate an mRNA [25-27]. Codon bias at the beginning of the translated mRNA sequence could in principle modulate the number of ribosomes that are sequestered by a mRNA if the rates of elongation at the first codons were sufficiently slow that stalled ribosomes could block access to the initiation signals. In order for such a mechanism to work in a systematic way, the initial coding sequences of genes would need to be correlated with the expression levels of the genes. However, Bulmer [5] has shown for *E. coli* and for *S. cerevisiae* that the initial coding sequences of highly expressed, intermediate and weakly expressed genes are all biased in similar ways. This means that regulation of expression levels via ribosome queueing at the beginning of the

mRNA sequences for these different groups of proteins is not feasible.

Nevertheless we know that an intragenic codon bias is relevant to expression levels from the results of Nassal et al. [19]. One clue to their function is provided by observations relating the efficiency of initiation to the function of sequences within the coding regions of genes [28]. In particular, it has been noted that there are sequences at the 5′ end of the 16S ribosomal RNA that are complementary to four or more adjacent nucleotides in the first 16 nucleotides of highly expressed mRNAs, while only three nucleotides are matched for low protein expression level mRNAs in *E. coli* [29]. In addition, sequences at the 3′ end of the 16S ribosomal RNA are complementary to between 6 and 12 nucleotides within the string from nucleotide 15 to 29 in coding sequences; here, the degree of matching with the consensus sequence is correlated with the efficiency of translation for those mRNA species for which such data are available [30].

Accordingly, there is reason to believe that interactions between ribosomal RNA and the initial coding sequences of mRNAs influence the efficiency of translational initiation. Nevertheless, the statistical uniformity of the codon frequencies observed in the initial coding sequences of mRNA translated at very different expression levels [5] suggests that something else must be involved in the selection of this intragenic bias. It may be that a uniform, optimal separation of ribosomes along the polysome is achieved by the initial codon strings. On the other hand, it might be that the virtues of this intragenic codon bias have nothing to do with translation.

4. CODON-SPECIFIC TRANSLATION RATES

There is general agreement that codons are translated at different rates. There is even some evidence to support this view, though such data are not overabundant. The first indication of non-uniform translation rates was the observation that there are pauses during polypeptide elongation and that these can be identified with short strings of rarely used codons [10,14,15]. More direct are the measurements of translation rates that show nearly a two-fold greater rate for mRNA species with predominantly common codons compared to those containing a greater frequency of rarely used codons [16]. Recently, the introduction of strings of common codons into the *lacZ* gene has revealed that these are translated at least six-fold faster on average than are strings of rarely used codons [17].

Three sorts of parameters can influence codon-specific translation rates. One is the maximum turnover rate (k_{cat}) of the ribosome which for a variety of reasons might vary from codon to codon. Another is the efficiency with wich aminoacyl-tRNA species are matched with the cognate codon on the ribosome. This function can be characterized by the rate factor (R) for the elongation factor Tu-aminoacyl-tRNA-GTP (ternary) complex-ribosome interaction as described by Fersht [31]. The R factor together with the third parameter, the ternary complex concentration determines the translation rate even when the codon programmed ribosome is not served by saturating concentrations of ternary complex.

According to tradition the bacterial ribosome has been thought to operate most of the time at its maximum rate; that is to say, it has been assumed that bacterial ribosomes are served by saturating concentrations of ternary complexes [32]. However, the evidence speaks rather clearly against this view. First, ribosomal mutants with reduced translation rates do not translate with a k_{cat} significantly lower than that of wild type ribosomes [33]. In contrast, such mutants do have significantly lower R factors that can be correlated with their lower elongation rates [34]. Second, mutant bacteria growing with lower concentrations of elongation factor Tu than wild-type, have correspondingly slower elongation rates (I. Tubelakas and D. Hughes, personal communication). Finally, natural isolates of *E. coli* vary very greatly in their growth rates as well as in their ribosome phenotypes; nevertheless, there is a very tight correlation between their R factors, translational elongation rates and growth rates [35]. All of these observations suggest that most of the codons in the mRNA species normally translated by *E. coli* are served by ternary complex concentrations that are well below saturation levels.

If the codon programmed ribosome is not in general kinetically saturated by ternary complex, we must expect that significant variations in the concentrations of individual tRNA species are reflected in variations in the rates of translation of the corresponding codons. The relevant observations are straightforward. First, for one particular codon, AGG, the apparent translation rate responds to variations of the cognate tRNA concentration in vivo [36]. More generally, there is a very clear correlation between a high abundance for a subgroup of tRNA species on the one hand and the relatively high frequency with which their cognate codons are used to code the amino acid sequences of the most abundant proteins in the bacteria [1,4,6,7,13,14,37–39]. It is precisely this subset of so-called major codons which is found to be translated at the fastest rates by Pedersen and his colleagues [16,17]. It seems rather clear that tRNA concentration is one of the important parameters determining the variation of translation rates at individual codons.

Other kinetic parameters may be relevant here. There are tRNA species that translate more than one codon and differences in the translation rates for such isocodons are expected to arise from kinetic differences in the mechanism of translation of the isocodons, for example, in the translocation rates on the ribosome. That the same tRNA may be matched with different

isocodons at somewhat different rates has been suggested by the data of Curran and Yarus [18]. However, this need not have a very great effect on the overall translation rate since codon matching may not be a very big part of the elongation cycle [40].

More provocative is the observation that the rate of translation by the sole Glu-tRNA isoacceptor at GAG is one third of that at AAA [41]. Clearly, for this pair of codons there is a significant kinetic difference somewhere in the elongation cycle. The problem with this observation is that it is unique, and that means that we cannot say yet how general or restricted such effects may be. Likewise, there are a few observations showing that the expression of synthetic or heterologous genes can be accompanied by unusually high codon-specific missense errors in the corresponding proteins [42] (C. Scorer, M. Carrier and R. Rosenberger, personal communication). These isolated observations are important because they emphasize that there may be codons and strings of codons that are avoided or selected because they have idiosyncratic effects on the elongation mechanism that are independent of the indirect effects of their corresponding tRNA species. Having said that, we will now concentrate on the indirect effects of tRNA species on the selection of codon sequences.

5. THE MAJOR CODON BIAS

So far, all the interpretations of codon bias that we have discussed are based on a particular view; namely, that of gene expression as seen from the perspective of individual genes and their products. Now we will shift to a more global perspective in which the overall efficiency of translation and its relationship to cell growth dominate our view.

We have already remarked on the fact that the expression level of a given gene product in general cannot be influenced by the rate of translation if its mRNA represents only a small fraction of the total mRNA pool. Nevertheless, the efficiency of the production of that particular protein is greater if the codon bias of the mRNA is such as to support rapid elongation rates rather than slow rates. Here, by efficiency we mean the rate of production of the protein normalized to the mass of the translation apparatus engaged in making it [43]. Again, if most of the different mRNA species that are being translated are made up of a biased subset of fast codons, the efficiency of production of all of these proteins will be greater than if the bias were different. Furthermore, if the abundance of the tRNA isoacceptor species is matched with the codon bias, the efficiency of translation will be correspondingly enhanced because this will minimize the mass of aminoacyl-tRNA-GTP-EFTu ternary complex that is employed to translate these mRNA species at any particular rate. These notions are the basis of our current view of the function of the major codon bias [20,44].

The basic idea is that in rich media translational efficiency is expected to be more critical to maximum growth rates than it is in poorer media. This expectation follows from the fact that in rich media a greater fraction of the metabolic activity is devoted to translation than it is in poor media [43]. This expectation has been verified by studying mutants with impaired translation kinetics under different growth conditions [45].

In addition, the protein composition of bacteria changes when they grow in different media: a small number of different proteins dominate in rich media, while a larger number of different proteins are expressed in lesser individual amounts in poor media [46,47]. This creates the opportunity to preferentially code the dominant group of proteins expressed at high growth rates by a very biased subset of codons. These would correspond to the major codon preference. The advantage of such an arrangement to the organism would be expressed as a gain in the efficiency of translation corresponding to a reduction in the total demand for ternary complex.

Thus, maximum efficiency of translation requires a maximum rate of translation normalized to a minimum mass of translation equipment. Relevant here is the requirement for a minimum mass of the ternary complex aminoacyl-tRNA-GTP-EFTu. At the fastest growth rates the problem is to raise the ternary complex concentration corresponding to the translated codons so that the maximum rates are approached and at the same time to minimize the total amounts of ternary complex. This can be done by matching the tRNA isoacceptor abundance to the biased codon frequencies of the major proteins [20,44]. In such an arrangement the increased concentrations of tRNA ternary complexes required to match the major codons would be compensated by a decreased concentration of the ternary complexes corresponding to the other tRNA species. The unique prediction of this interpretation is that tRNA abundance changes in predictable ways when the growth conditions are changed; the richer the medium, the higher will be the concentration of tRNA species that translate the major codons and the lower will be the concentration of the remainder of isoacceptor species.

This prediction has not yet been tested exhaustively by measuring the growth rate dependence of the abundance for all of the tRNA species in a suitable organism. However, nearly half of the isoacceptors in *E. coli* have been studied. Thus, 19 species corresponding to the isoacceptors for Leu and Met [48], as well as those for Arg, Gly and Pro (Emilsson and Kurland, unpublished data) have been analyzed. A very gratifying agreement with expectations has been observed. Large variations in the relative amounts of tRNA species that translate major and minor codons are observed in the expected directions.

In summary, the unique character of the predictions and the degree of agreement with the observations are

persuasive. It seems that the major codon bias is nothing less than an arrangement that permits the cell to maximize the efficiency of translation at the fastest growth rates. Here, a special subset of proteins, coded by a correspondingly biased subset of codons can be translated by a suitably biased tRNA population in order to minimize the tRNA ternary complex mass required to maintain fast translation.

Acknowledgements: My work is supported by grants from the Swedish Cancer Society and Natural Sciences Research Council. I am grateful to V. Emilsson, D. Hughes, S. Pederssen and R. Rosenberger for helpful discussions and for sharing unpublished data with me.

REFEFENCES

[1] Grantham, R., Gautier, C. and Gouy, M. (1980) Nucleic Acids Res. 8, 1893-1911.
[2] Bernardi, G. and Bernardi, G. (1985) J. Mol. Evol. 22, 363-365.
[3] Sharp, P.M., Shields, D.C., Wolfe, K.H. and Li, W.-H. (1989) Science 246, 808-810.
[4] Gouy, M. and Gautier, C. (1982) Nucleic Acids Res. 10, 7055-7074.
[5] Bulmer, M. (1988) J. Theor. Biol. 133, 67-71.
[6] Chavancy, G. and Garel, J.-P. (1981) Biochimie 63, 187-195.
[7] Grosjean, H. and Fiers, W. (1982) Gene 18, 199-209.
[8] Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucleic Acids Res. 12, 6663-6671.
[9] Bonekamp, F., Dalbøge Andersen, H., Christensen, T. and Jensen, K.F. (1985) Nucleic Acids Res. 13, 4113-4123.
[10] Varenne, S. and Lazdunski, C. (1986) J. Theor. Biol. 120, 99-110.
[11] Garel, J.-P. (1974) J. Theor. Biol. 43, 211-225.
[12] Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
[13] Ikemura, T. (1981) J. Mol. Biol. 151, 389-409.
[14] Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) J. Mol. Biol. 180, 549-576.
[15] Randall, L.L., Josefsson, L.-G. and Hardy, S.J.S. (1980) Eur. J. Biochem. 107, 375-379.
[16] Pedersen, S. (1984) EMBO J. 3, 2895-2898.
[17] Sørensen, M.A., Kurland, C.G. and Pedersen, S. (1989) J. Mol. Biol. 207, 365-377.
[18] Curran, J.F. and Yarus, M. (1989) J. Mol. Biol. 209, 65-77.
[19] Nassal, M., Mogi, T., Karnik, S.S. and Khorana, H.G. (1987) J. Biol. Chem. 262, 9264-9270.
[20] Andersson, S.G.E. and Kurland, C.G. (1990) Microbiol. Rev. 54, 198-210.
[21] Chaney, W.G. and Morris, A.J. (1979) Arch. Biochem. Biophys. 194, 283-291.
[22] Yamamoto, T., Suyama, A., Mori, N., Yokota, T. and Wada, A. (1985) FEBS Lett. 181, 377-380.
[23] Gabain, A.V., Belasco, J.G., Schottel, J.L., Chang, A.C.Y. and Cohen, S.N. (1983) Proc. Natl. Acad. Sci. USA 80, 653-657.
[24] Petersen, C. (1987) Mol. Gen. Genet. 209, 179-487.
[25] Stanssens, P., Remaut, E. and Fiers, W. (1986) Cell 44, 711-718.
[26] Hoekema, A., Kastelein, R.A., Vasser, M., de Boer, H.A. (1987) Mol. Cell. Biol. 7, 2914-2924.
[27] Liljenström, H. and v. Heijne, G. (1987) J. Theor. Biol. 124, 43-55.
[28] Gold, L. (1988) Annu. Rev. Biochem. 57, 199-233.
[29] Petersen, G.B., Stockwell, P.A. and Hill, D.F. (1988) EMBO J. 7, 3957-3962.
[30] Sprengart, M.L., Fatscher, H.P. and Fuchs, E. (1990) Nucleic Acids Res. 17, r1-r172.
[31] Fersht, A.R. (1977) Biochemistry 16, 1025.
[32] Maaløe, O. (1979) in: Biological Regulation and Development, (R.F. Goldberger, Ed.) Plenum, New York, pp. 487-542.
[33] Kurland, C.G. and Ehrenberg, M. (1987) Annu. Rev. Biophys. Biophys. Chem. 16, 291-317.
[34] Andersson, D.I., v. Verseveld, H.W., Stouthamer, A.H. and Kurland, C.G. (1986) Arch. Microbiol. 144, 96-101.
[35] Mikkola, R. and Kurland, C.G. (1991), in press.
[36] Mistra, R. and Reeves, P. (1985) Eur. J. Biochem. 152, 151-155.
[37] Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) Nucleic Acids Res. 8, r49-r53.
[38] Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
[39] Ikemura, T. and Ozeki, H. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, 1087-1097.
[40] Bilgin, N., Kirsebom, L.A., Ehrenberg, M. and Kurland, C.G. (1988) Biochimie 70, 611-618.
[41] Sørensen, M.A. and Pedersen, S. (1991) J. Mol. Biol. in press.
[42] Bogosian, G., Violand, B.N., Jung, P.E. and Kane, J.F. (1990) in: The Ribosome: Structure, Function and Evolution (A.D.W.E. Hill, R.A. Garrett, P.B. Moore, D. Schlessinger and J.R. Warner, eds.) American Society for Microbiology, Washington.
[43] Ehrenberg, M. and Kurland, C.G. (1984) Q. Rev. Biophys. 17, 45-82.
[44] Kurland, C.G. (1987) Trends Biochem. Sci. 12, 126-128.
[45] Mikkola, R. and Kurland, C.G. (1988) FEMS Microbiol. Lett. 56, 265-270.
[46] Pedersen, S., Bloch, P.L., Keeh, S. and Neidhardt, F.C. (1978) Cell 14, 179-190.
[47] Emilsson, V. and Kurland, C.G. (1990) EMBO J. 9, 4359-4366.
[48] Emilsson, V. and Kurland, C.G. (1990) Biochim. Biophys. Acta 1050, 248-251.