Third letters in codons counterbalance the (G+C)-content of their first and second letters

Akiyoshi Wada and Akira Suyama

Department of Physics, Faculty of Science, University of Tokyo, Bunkyo-ku, Tokyo, Japan

Received 8 July 1985

The correlation among the frequencies of appearance of G and C bases at the first, second and third positions of codons is examined. It is found that the redundancy at the wobbling third base is utilized to counterbalance the local (G+C)-content variation due to the first and second bases of codons, so that a homogeneous (G+C)-content is provided to each individual gene. We speculate that this levelling tendency of (G+C)-content comes from the functional constraint to achieve a uniform double-helix stability in a gene.

Double helix stability *Codon usage* *Gene homostabilizing propensity* *DNA* *Codon*

1. INTRODUCTION

Analyses of the correlation between the double-helix stability distribution and the location of genes in a number of DNAs have disclosed the tendency for each individual gene to have a uniform stability distribution around its own characteristic stability value [1–6]. We call this the gene homostabilizing propensity. Here, we examined whether the third letter of codons, which is the wobbling site, plays an essential role in producing a homogeneous (G+C)-content distribution, and thus a homogeneous stability, in individual genes. We studied the correlation among the frequencies of appearance of G and C bases at the first, second, and third positions of codons. The average local (G+C)-content at the first plus second sites in a gene has a remarkable negative correlation with that of the third site. In other words, each gene realizes a homogeneous (G+C)-content distribution by utilizing the redundancy at the wobbled third base to counterbalance the (G+C)-content variation at the first and second bases of codons.

2. RESULTS

The protein-coded gene is a sequence of codons, that is, $X_1Y_1Z_1X_2Y_2Z_2 \ldots X_iY_iZ_i \ldots X_NY_NZ_N$, where X, Y and Z denote the first, second, and third bases of codons, respectively; subscript numbers indicate codon numbers and N is the number of codons in a gene. First, we calculated (G+C)-contents of these sites locally averaged over a short segment ($2\Delta + 1$ codons from $i - \Delta$ to $i + \Delta$) in a gene:

$$\overline{(GC)}_{X_i}^{\text{local}} = \sum_{i-\Delta}^{i+\Delta} (GC)_{X_i}/(2\Delta + 1),$$

where X can be Y or Z, and $(GC)_{X_j} = 1$ if the X_j site is occupied by G or C, and 0 otherwise. The same equations were used for Y_i and Z_i sites. Next, in order to inspect whether the correlation is an intragene or an intergene event, we calculated the (G+C)-content of X, Y, and Z averaged over the whole length of a gene:

$$\overline{(GC)}_X^{\text{gene}} = \sum_{i=1}^{N} (GC)_{X_i}/N,$$

where X can be Y or Z.

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

The diagrams in fig.1 display typical correlation plots between $\overline{(G+C)}^{local}_{X+Y}$ and $\overline{(G+C)}^{local}_{Z}$ in each of several genes. The base sequences of genes are drawn from Gen Bank release 29. Each point indicates one of the above-mentioned values, which are averaged over a segment of 63 bases (21 codons, i.e. $\Delta = 10$). This averaging window is moved from the 5′- to the 3′-terminus of a gene, so that the distribution of the points indicates the nature of the correlation from $i = \Delta + 1$ to $N - \Delta - 1$, and the correlation coefficients, r, are obtained as indicated in the upper portion of the diagrams.

Negative correlations are clearly exhibited in each case. This kind of profile is generally common to each gene examined, while no meaningful correlation is found between X and Y letters. The results of the correlation study extended over many prokaryote genes are summarized in the form of the histogram of correlation coefficients in fig.2. The histogram displays the number of genes (ordinate) which have the correlation coefficient indicated on the abscissa. The 161 genes used in these statistics are from following species: Phages ϕX174, G4, fd, T7, and lambda$^+$; *E. coli* ribosomal proteins S8au L6 genes*; Mycoplasma ribosomal proteins S8 and L6 genes*; *Staphylococcus aureus* plasmid pT181$^+$; and *S. aureus* plasmid pC194$^+$. ($^+$ Gen Bank release 29, * A. Muto et al. (1984) Nucleic Acid Res. 12, 8209–8217). This histogram clearly demonstrates that the negative correlation is a general trend in the codon. In other words, the third sites of codons in each gene act to counterbalance the change in (G + C)-content at their first and second sites. Similar examinations made for each of 3 different (0 base, 1 base and 2 base shifted) frames in the base sequence of non-protein coding regions, however, showed no such negative correlation as the protein coding regions revealed; the counterbalancing phenomenon is thus a characteristic feature of protein coding regions.

By contrast, $\overline{(GC)}^{gene}_{X+Y}$ and $\overline{(GC)}^{gene}_{Z}$ are found to have a positive correlation (fig.3). Differing from the diagrams in fig.1, each point in fig.3 indicates the average (G + C)-content over an entire stretch of each gene, so that the distribution of the points in one diagram shows the correlation including (G + C)-rich genes and (A + T)-rich genes. It is apparent that the first and second sites of

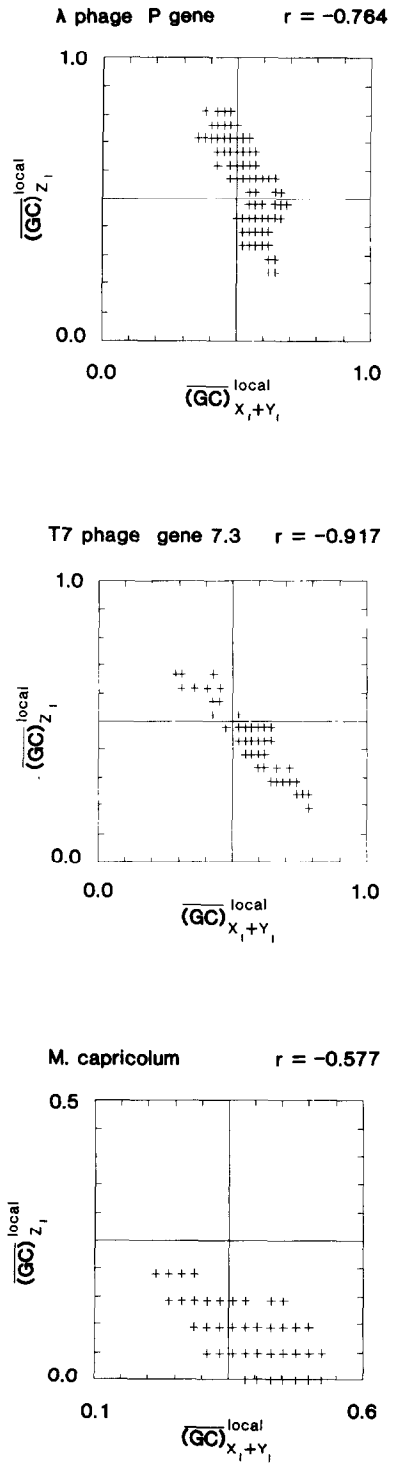

Fig.1. Typical correlation diagrams between $\overline{(GC)}^{local}_{X_i+Y_i}$ and $\overline{(GC)}^{local}_{Z_i}$ in 3 genes giving different (G + C)-content.

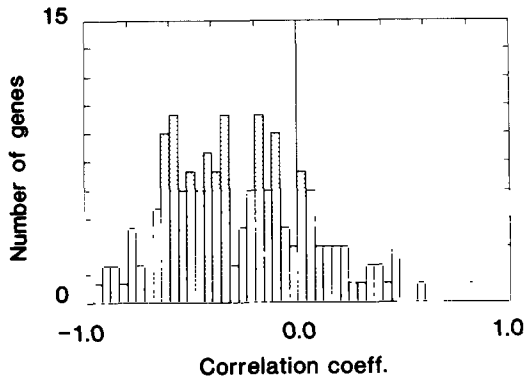

Fig.2. Histogram showing the distribution of the correlation coefficients, a few samples of which are shown in fig.1.

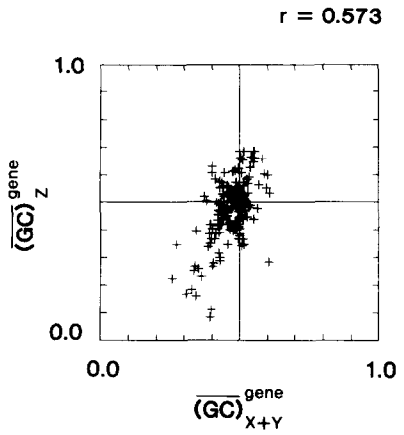

Fig.3. Correlation diagram between $\overline{(G+C)}_{X+Y}^{gene}$ and $\overline{(G+C)}_Z^{gene}$.

codons contribute little to the change in the (G+C)-content of a gene, while the (G+C)-content of the third site changes greatly from 0.08 to 0.7. The diagram is in striking contrast to fig.1 as it shows a weak positive correlation between X+Y and Z sites.

3. DISCUSSION

It is well known that synonymous codons are used in a non-random manner in both prokaryotes and eukaryotes [7–14]. This non-random use of codons is believed to come from constraints that arise from the need to realize a high accuracy of translation and/or optimal efficiency of protein synthesis.

Nomura and his colleagues [8,9] found that the codons in *E. coli* ribosomal protein genes prefer those codons recognized by the most abundant t-RNAs in *E. coli* cells. Ikemura showed by examining a number of codons in *E. coli* [10,11] and yeast [12] that this is the general trend of codon choice. He studied the constraints of codon usage in these organisms from the viewpoint of translational efficiency. In most of the genes, a strong positive correlation between molar quantity of t-RNA and the frequency of usage of respective codons was found. Namely, genes encoding abundant proteins selectively used codons of dominant t-RNA, while completely neglecting those of t-RNAs of small quantity. Therefore, the codon choice in *E. coli* and yeast genes, and probably in general in living organisms, seems to be constrained by t-RNA, its availability and by translational efficiency.

Grosjean et al. [13,14] proposed the hypothesis that an efficient translation is facilitated by proper choice of degenerate codewords promoting a codon-anticodon interaction with intermediate strength, avoiding those with very strong or very weak interaction energy. They demonstrated that this rule of the codon usage is realized in MS2 RNA bacteriophage and in mRNAs of *E. coli* [13,14].

The counterbalancing effect of the third letter of codons, which was found in this study, may be explained in one of two ways: (i) the third letter plays a role in providing a homogeneous (G+C)-content distribution in each gene. The internal homology in the double-helix stability of each gene thus produced will provide the functional merit in genetic events such as the smooth zipper-opening of DNA double helix at the duplication process. Or, (ii) the genetic coding system prefers, in general, the medium codon-anticodon interaction strength, as Grosjean et al. proposed.

Both of these cases may, to a greater or lesser degree, provide constraint to the third letter selections. The present authors, however, favor the first factor for the following reasons. If the genetic system avoided codons of extreme (G+C)- or (A+T)-content, then genes of high or low (G+C)-content would have been eliminated; this is contrary to the fact. Furthermore, the positive correlation between $\overline{(G+C)}_{X+Y}^{gene}$ and $\overline{(G+C)}_Z^{gene}$ exhibited

in fig.3 is not explained by the second hypothesis.

Sueoka [15] and Freese [16] proposed a species-specific balance in the equilibrium of the GC AT transversion, so that each species, and probably genes also, tend to have their own characteristic values of (G + C)-content. The positive correlation exhibited in fig.3 indicates that all 3 letters in a codon cooperate to change a gene's (G + C)-content and to attain its characteristic value.

We consider that the role which Grosjean et al. [13,14] proposed is applicable for the smooth and correct reading of a set of codons in a gene of not only medium (G + C)-content but also of high or low (G + C)-content. Namely, to maintain both high efficiency and high accuracy of the translation, i.e. codon-anticodon interaction, all codons in a gene should have equivalent codon-anticodon interaction energy. This functional constraint may be one of the origins of the gene homostabilizing propensity.

Summarizing this study: (i) in each gene the third site of a codon which has a wobbling characteristic counterbalances the (G + C)-content at the first and second sites so that (G + C)-content distribution in the gene is homogeneous. (ii) This counterbalancing nature does not exist in any of the 3 possible frames in the base sequence of non-protein coding regions. (iii) The average (G + C)-content of each gene is altered mainly by base substitution at the third site of the codon and (iv) from the positive correlation between the X + Y and Z sites, which is revealed when they are averaged over the entire length of a gene, every base in a codon seems to work cooperatively toward realizing the gene's characteristic value of (G + C)-content.

Finally, we speculate that this smoothing tendency of (G + C)-content in a gene comes from the need to have a uniform double-helix stability and/or homogeneous codon-anticodon interactions in the gene [1–6]. This constraint, even though very weak, seems to have created the described base sequence uniformity during the period of biological evolution.

ACKNOWLEDGEMENTS

We thank Dr T. Ikemura for helpful discussions. We also thank Dr H. Ozeki and S. Osawa for their encouragement. This work was supported by a grant-in-aid from the Ministry of Education, Science and Culture, Japan.

REFERENCES

[1] Wada, A., Tachibana, H., Gotoh, O. and Takanashi, M. (1976) Nature 263, 439–440.
[2] Wada, A., Yabuki, S. and Hushimi, Y. (1979) CRC Crit. Rev. Biochem. 9, 87–144.
[3] Suyama, A. and Wada, A. (1983) J. Theor. Biol. 105, 133–145.
[4] Wada, A. and Suyama, A. (1983) J. Phys. Soc. Jap. 52, 4417–4422.
[5] Wada, A. and Suyama, A. (1984) J. Biomol. Structure Stereodyn. 2, 573–591.
[6] Wada, A. and Suyama, A. (1985) in: The Molecular Bases of Cancer: An Interdisciplinary Discussion on Basic and Applied Aspects of Cancer (Rein, R. ed.) Alan R. Liss Inc., New York.
[7] Grantham, R. (1980) Trends Biochem. Sci. 5, 327–331 and references therein.
[8] Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. and Dennis, P.P. (1979) Proc. Natl. Acad. Sci. USA 76, 1697–1701.
[9] Nomura, M., Post, L.E. and Jinks, S. (1980) in: RNA Polymerase, t-RNA and Ribosomes (Osawa, S. ed.) pp.315–328, Univ. of Tokyo Press, Japan.
[10] Ikemura, T. (1981) J. Mol. Biol. 146, 1–21.
[11] Ikemura, T. (1981) J. Mol. Biol. 151, 389–409.
[12] Ikemura, T. (1982) J. Mol. Biol. 158, 573–579.
[13] Grosjean, H., Sankoff, D., Min Jou, W., Fiers, W. and Cedergen, R.J. (1978) J. Mol. Biol. 12, 113–119.
[14] Grosjean, H. and Fiers, W. (1982) Gene 18, 199–209.
[15] Sueoka, N. (1962) Proc. Natl. Acad. Sci. USA 48, 582–592.
[16] Freese, E. (1962) J. Theor. Biol. 3, 82–89.