Volume 7 Number 8 1979 Nucleic Acids Research

The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other

Takashi Miyata, Hidenori Hayashida, Teruo Yasunaga and Masami Hasegawa*

Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, and *The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minatoku, Tokyo 106, Japan

Received 29 August 1979

ABSTRACT

The pattern of codon utilization in the variable and constant regions of immunoglobulin genes are compared. It is shown that, in these regions, codon utilizations are quite distinct from one another: For most degenerate codons, there is a selective bias that prefers C and/or G ending codons to U and/or A ending codons in the constant region compared with the bias in the variable region. This would strongly suggest that, in immunoglobulin genes, the bias in code word usage is determined by other factors than those concerning with the translational mechanism such as tRNA availability and codon-anticodon interaction. A possibility is also suggested that this difference of code word usage between them is due to the existence of secondary structure in the constant region but not in the variable region.

INTRODUCTION

The comprehensive comparison of different kinds of genes has revealed that there exist distinct biases in the use of degenerate codons (see e.g., ref. 1). Several arguments have been noted on the primary factors determining the bias of codon utilization: They are 1) tRNA availability[1-3], 2) codon-anticodon interaction[4, 5], 3) selective pressure operating on the base content of DNA[6], 4) secondary structure of mRNA[7, 8] and 5) regulatory mechanism and other factors[9, 10]. Most genes show complicated patterns of codon utilization, which may be determined by many factors. However, relatively simple patterns of codon usage are observed in certain genes like ATPase genes of yeast mtDNA[6], ribosomal protein genes of E.coli[3] and MS2 phage RNA[8, 16-18], which may well be explained by only one or a few factors. These genes may be of particular interest, because they provide much insights about the typical pattern of codon utilization being determined by a single factor, and they also make feasible the analysis of which factor is a primary determinant among the possible factors in a certain gene. Here we show that, in the variable (V) and constant (C) regions of immunoglobulin genes, codon utilizations are quite different from one another. A possibility is

also suggested that this difference of code word usage between them is due to the existence of secondary structure in C region but not in V region.

METHOD

The frequency of codons and the composition of amino acids vary with genes that code for them. For a quantitative comparison of patterns of codon usage between different genes or regions of a gene, it might therefore be appropriate to use an index f_α , representing the frequency of codon α normalized among the frequencies of degenerate codons specifying the same amino acid A:

$$f_\alpha = n_\alpha / \sum_{\alpha \epsilon A} n_\alpha \quad , \qquad (1)$$

where n_α stands for the frequency of codon α. Observing the frequencies of codons in V and C regions (i.e., n_α(V-region) and n_α(C-region), α=1,2,...,64) separately, and estimating f_α(V-region) and f_α(C-region) for all the codons, a quantitative analysis is possible for the degree of difference of preferential codon usages between the two regions for every codons. For statistically valid comparison to be made between V and C regions, data for n_α(V-region) are accumulated from two λ[11] and two κ[12] light chain V regions, and data for n_α(C-region) are from κ light chain C region[13], 5'-terminal sequence of λ light chain C region[11] and fragments of γ_1 heavy chain C region[14], for which the nucleotide sequences are now available. Table 1 shows the frequencies of codons in V and C regions, and also the frequency of codons whose third positions are in the base pairing sites of MS2 RNA[8] for comparison(see below).

Most degenerate codons differ from each other only in the third position of codons except six-fold degenerate codon groups, for which we tentatively divided into two-fold and four-fold degenerate codon groups as a matter of convenience; i.e., Leu_2(UUR) and Leu_4(CUX) for leucine, Arg_2(AGR) and Arg_4(CGX) for arginine, and Ser_2(AGY) and Ser_4(UCX) for serine, where X stands for any one of four nucleotides, and Y and R for pyrimidine and purine, respectively. According to this classification, selective biases in the use of synonymous codons can be analysed in terms of the composition of nucleotides at the third position of codons.

RESULTS AND DISCUSSION

Figure 1 shows a plot of f(V-region) (Y axis) *versus* f(C-region) (X axis) for all the codons except non-degenerate codons (i.e., methionine

Table 1. The frequencies of codons in the variable (V) and the constant (C) regions of immunoglobulin genes, and the frequency of codons whose third positions are in the base-pairing sites (PS) of MS2 phage RNA.

		V	C	PS			V	C	PS			V	C	PS			V	C	PS
Phe	UUU	4	2	8	Ser	UCU	9	5	7	Tyr	UAU	11	3	5	Cys	UGU	8	5	4
	UUC	9	6	21		UCC	5	4	19		UAC	7	4	22		UGC	0	4	4
Leu	UUA	4	1	11		UCA	10	4	8	Stp	UAA	-	-	-	Stp	UGA	-	-	-
	UUG	4	2	8		UCG	0	1	20		UAG	-	-	-	Trp	UGG	8	3	17
	CUU	0	0	8		CCU	7	5	12	His	CAU	6	1	5		CGU	0	0	10
Leu	CUC	8	3	21	Pro	CCC	5	8	7		CAC	1	2	8	Arg	CGC	2	0	15
	CUA	7	0	9		CCA	8	10	6	Gln	CAA	6	1	7		CGA	4	2	7
	CUG	3	3	9		CCG	2	2	12		CAG	13	5	17		CGG	0	1	11
	AUU	4	4	8		ACU	17	6	10	Asn	AAU	4	3	9	Ser	AGU	16	4	4
Ile	AUC	8	5	18	Thr	ACC	10	11	16		AAC	15	8	17		AGC	9	10	13
	AUA	4	1	10		ACA	16	6	7	Lys	AAA	3	8	9	Arg	AGA	4	1	4
Met	AUG	3	2	15		ACG	0	3	12		AAG	4	14	20		AGG	4	3	6
	GUU	7	1	10		GCU	12	4	16	Asp	GAU	9	6	17		GGU	12	5	21
Val	GUC	8	8	18	Ala	GCC	7	8	18		GAC	4	6	16	Gly	GGC	4	5	13
	GUA	1	2	12		GCA	15	2	11	Glu	GAA	10	3	7		GGA	9	3	5
	GUG	6	8	11		GCG	0	0	19		GAG	5	11	25		GGG	7	0	13

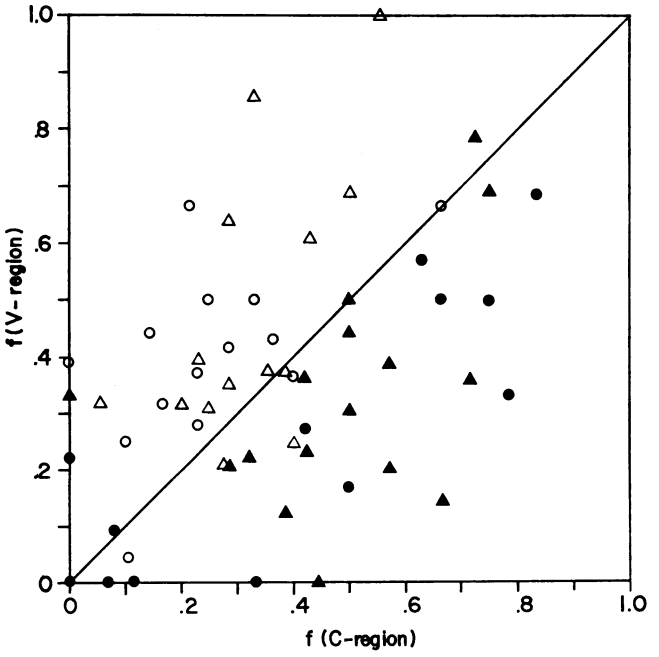

Fig. 1. A plot of f(V-region) (Y axis) versus f(C-region) (X axis). Points Δ, ▲, o and • correspond to U, C, A and G ending codons, respectively. Solid line shows Y=X.

and tryptophan). Points in this figure are distinguished from each other according to the nucleotide type in the third position of codons. Most points corresponding to the codons ending C or G are found in a domain $Y < X$, and conversely, points for U or A are found in the other domain $Y > X$. That is, there is a selective code word usage that prefers C and/or G to U and/or A in the third position of degenerate codons in C region compared with the bias in V region. This difference in codon utilization is statistically significant: If there is no difference in the biases of codon utilization between V and C regions, it is expected that the probability that a point corresponding to C or G is found in the domain $Y < X$ is 1/2. As the probability that, out of the n points corresponding to C or G, r points are found in $Y < X$ follows a binomial distribution, then, $t \left(= (r - n\times(1/2))/\sqrt{n\times(1/2)\times(1/2)} \right)$ follows a normal distribution with mean being equal to zero and variance being equal to unity. Observing n=29 and r=23 from this figure (the points on the line $Y = X$ are excluded from r), we have t=3.16. The probability that this value is realized is smaller than 10^{-3}. We therefore conclude that the preferential codon usages in the variable and constant regions of immunoglobulin genes are quite distinct from one another. This finding would suggest that in the immunoglobulin gene, the bias in code word utilization is determined by other factors than those concerned with the translational mechanism such as tRNA availability and codon-anticodon interaction, because, though V and C regions lie separate from each other in DNA and precursor mRNA, they come together to form a single translational unit in mature mRNA[15].

It should be explained why the patterns of codon usage differ between the variable and constant regions of immunoglobulin gene. Recently, Hensgens et al.[6] have shown that, in ATPase gene of yeast mtDNA, the codon usage is biased in favor of A+U rich codons, and have suggested that this is due to a selective pressure towards a low G+C content operating both on the non-coding and coding sequences. That is, there is a selective bias favoring the use of characteristic nucleotide content on the DNA level. If this is really a case for immunoglobulin gene, it is expected that the nucleotide contents at the third position of codons in V and C regions correlate well with the nucleotide contents of non-coding sequences around the respective regions, and the difference of codon utilization can be explained on the basis of the distinct selective biases favoring the use of characteristic nucleotide contents in the non-coding sequences of respective regions. We estimate the nucleotide contents for V and C regions from the sense strand of non-coding

segments within and surrounding the corresponding regions. From these contents, the expected frequencies of bases are estimated every regions, and goodness of fit test can now be made (Table 2(a)). As the table shows, the hypothesis is rejected at a significant level. This implies that this selective pressure is not a primary factor resulting in the difference of codon usage between V and C regions.

We also attempt to analyse a possibility that this difference depends on whether an extensive secondary structure exists or not in the respective regions. Fiers and his coworkers[16-18] have shown that MS2 phage involves an extensive secondary structure within its RNA. Previously, we have shown on the basis of the secondary structure proposed by Fiers et al. that, in base-pairing sites of MS2 RNA, there is a selective bias favoring C and/or G over U and/or A at the third position of codons and *vice versa* in the non-pairing sites, which is interpreted as a result of selective constraint that stabilizes the secondary structure of the RNA[8]. If a certain gene involves a secondary structure within its mRNA, it is expected that the gene shows a pattern of codon usage similar to that in pairing sites of MS2 RNA. According to our previous analysis[8], nucleotide contents at the third position of

Table 2. Distributions of U, C, A and G observed at the third position of degenerate codons in V and C regions of immunoglobulin genes and expected from (a) the frequencies in the non-coding regions of the genes and (b) the frequencies at the third position of codons in pairing sites of MS2 phage.

	U	*C*	*A*	*G*	*Goodness of fit test*
(a) Expected: Non-coding regions of immunoglobulin genes					
Observed(V-region)	*126*	*102*	*101*	*48*	$\chi^2 = 11.19$
Expected	*124.41*	*82.94*	*99.91*	*69.75*	*Pr. < 0.025*
Observed(C-region)	*54*	*92*	*44*	*53*	$\chi^2 = 15.02$
Expected	*74.60*	*73.63*	*53.22*	*41.55*	*Pr. < 0.005*
(b) Expected: Pairing sites of MS2 RNA					
Observed(V-region)	*126*	*102*	*101*	*48*	$\chi^2 = 98.57$
Expected	*79.92*	*127.43*	*58.44*	*111.22*	*Pr. < 0.001*
Observed(C-region)	*54*	*92*	*44*	*53*	$\chi^2 = 7.24$
Expected	*51.52*	*82.13*	*37.67*	*71.69*	*Pr. > 0.05*

The nucleotide contents in the non-coding regions of V(C) are estimated from the non-coding sequences within and surrounding the coding sequences of V(C), which are U=33.0%(30.7%), C=22.0%(30.3%), A=26.5%(21.9%) and G=18.5%(17.1%), respectively. The nucleotide contents at the third position of codons in the pairing sites of MS2 are U=21.2%, C=33.8%, A=15.5% and G=29.5%, respectively (ref. 8).

codons in pairing sites of MS2 RNA are U=21.2%, C=33.8%, A=15.5% and G=29.5%, respectively. Assuming that there exist secondary structures both in V and C regions, we have the expected frequencies shown in Table 2(b) for the respective regions. A goodness of fit test shows that, for V region, the hypothesis is rejected at a significant level, but is not rejected for C region. That is, the secondary structure might not be a primary factor by which the overall pattern of codon utilization in V region can be explained. However, a possibility can still not be excluded for C region. We therefore make a more detailed comparison between the pattern of codon usage in C region and that in the pairing sites of MS2 RNA.

We have already estimated the frequency of codons (i.e., n_α ,α=1,2,..., 64) whose third positions are in the pairing sites of MS2 RNA[8] (see Table 1). Applying this data to Eq. 1, the fraction f_α for the pairing sites of MS2 RNA (f_α(PS of MS2)) can be obtained. A comparison is thus possible for codon utilization between C region and the pairing sites of MS2 RNA. Figure 2 shows a plot of f(C-region) (Y axis) *versus* f(PS of MS2) (X axis) for all the degenerate codons (points corresponding to arginine codons CGX are excluded from the comparison due to small sample size (total number = 3) in C region). An appreciable correlation is found between f(C-region) and f(PS of

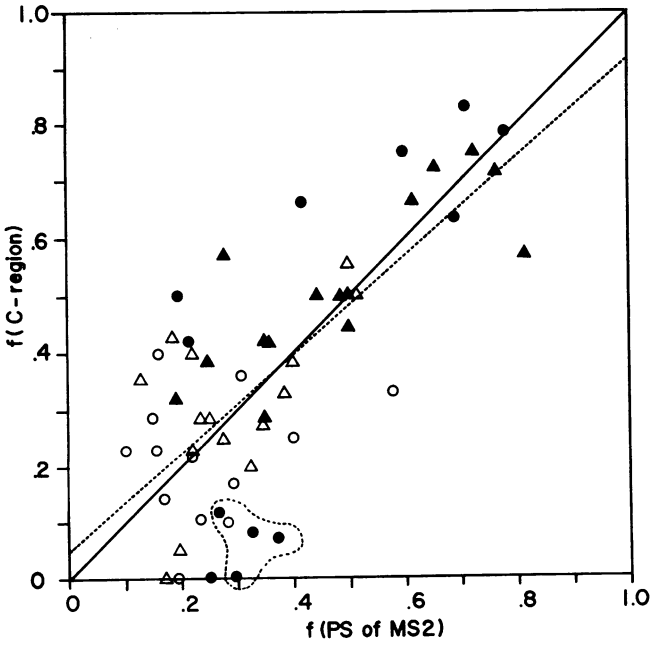

Fig. 2. A plot of f(C-region) (Y axis) versus f(PS of MS2) (X axis). Points Δ, ▲, o and • correspond to U, C, A and G ending codons, respectively. Solid line represents Y=X. Dotted line is the regression of Y on X: Y = 0.87X + 0.05, and γ (correlation coefficient) is 0.73. The points • in a dotted area correspond to codons involving CpG doublet.

MS2): Most points are found to be close to Y=X, and points corresponding to C and G ending codons are always larger than points corresponding to U and A except some limited number of points, especially points corresponding to codons involving CpG doublet which is infrequent to be found in eukaryotic mRNA (see e.g., ref. 1). Indeed, the correlation coefficient is 0.733 and the regression of Y on X is Y = 0.87X + 0.05, being close to Y = X. This would imply that there is an extensive secondary structure in C region. Thus the conclusion must be that the observed difference of codon utilizations between V and C regions mainly results from a secondary structure that would be expected to exist in C but not in V region. It should be noted that we do not exclude other possibilities, but we only suggest a primary factor by which most of the difference found in codon usage of V and C regions are explained.

Recently, Rogers et al.[19] have determined the nucleotide sequence encoding the C-terminal $1^1/_2$ domains of γ_1 constant region, and have compared the pattern of codon usage of the gene with those of other animal genes. On the basis of the overall nucleotide content in the third codon positions, they have grouped the patterns of codon usage into three classes. According to their classification, C and V regions belong to different groups from each other. They have proposed a possible secondary structure of γ_1 mRNA fragment.

ACKNOWLEDGEMENT We thank Professor H.Matsuda for continuous encouragement.

REFERENCES

1. Grantham,R. (1978) FEBS Letters, 95, 1-11.
2. Berger,E.M. (1978) J.Mol.Evol., 10, 319-323.
3. Post,L.E., Strycharz,G.D., Nomura,M., Lewis,H. & Dennis,P.P. (1979) Proc.Natl.Acad.Sci.USA, 76, 1697-1701.
4. Fitch,W.M. (1976) Science, 194, 1173-1174.
5. Grosjean,H., Sankoff,D., Min Jou,W., Fiers,W. & Cedergren,R.J. (1978) J.Mol.Evol., 12, 113-119.
6. Hensgens,L.A.M., Grivell,L.A., Borst,P. & Bos,J.L. (1979) Proc.Natl.Acad. Sci.USA, 76, 1663-1667.
7. Fitch,W.M. (1974) J.Mol.Evol., 3, 279-291.
8. Hasegawa,M., Yasunaga,T. & Miyata,T. (1979) Nucleic Acids Res., in press.
9. Fiers,W. & Grosjean,H. (1979) Nature, 277, 328.
10. Jukes,T.H. (1978) J.Mol.Evol., 11, 121-127.
11. Bernard,O., Hozumi,N. & Tonegawa,S. (1978) Cell, 15, 1133-1144.
12. Seidman,J.G.,Leder,A., Edgell,M.H., Polsky,F., Tilghman,S.M., Tiemeier, D.C. & Leder,P. (1978) Proc.Natl.Acad.Sci.USA, 75, 3881-3885.
13. Hamlyn,P.H., Brownlee,G.G., Cheng,C., Gait,M.J. & Milstein,C. (1978) Cell, 15, 1067-1075.
14. Sakano,H., Rogers,J.H., Hüppi,K., Brack,C., Traunecker,A., Maki,R., Wall, R. & Tonegawa,S. (1979) Nature, 277, 627-633.
15. Brack,C.,Hirama,M.,Lenhard-Schuller,R. & Tonegawa,S.(1978) Cell, 15,1-14.
16. Min Jou,W., Haegeman,G., Yaebaert,M. & Fiers,W. (1972) Nature, 237,82-88.
17. Fiers,W., Contreras,R., Duerinck,F., Haegeman,G., Merregaert,J., Min Jou,

W., Raeymarkers,A., Volckaert,G., Yaebaert,M., Van de Kerckhove,J., Nolf, F. & Van Montagu,M. (1975) Nature, 256, 273-278.
18. Fiers,W., Contreras,R., Duerinck,F., Haegeman,G., Iserentant,D., Merregaert,J., Min Jou,W., Molemans,F., Raeymaekers,A., Van den Berghe, A., Volckaert,G. & Yaebaert,M. (1976) Nature, 260, 500-507.
19. Rogers,J., Clarke,P. & Salser,W. (1979) Nucleic Acids Res., 6, 3305-3321.