Secondary structure of MS2 phage RNA and bias in code word usage

Masami Hasegawa*[†], Teruo Yasunaga** and Takashi Miyata**

*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106 Japan and

**Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812 Japan

Received 14 August 1979

ABSTRACT

Based on the secondary structural model of MS2 RNA, it is shown that, in base-pairing regions of the RNA, there is a bias in the use of synonymous codons which favours C and/or G over U and/or A in the third codon positions, and that in non-pairing regions, there is an opposite bias which favours U and/or A over C and/or G. This nature is interpreted as a result of selective constraint which stabilises the secondary structure of the single-stranded RNA genome of the MS2 phage.

INTRODUCTION

Recent accumulation of nucleotide sequence data for complete genomes (1-7) and for mRNAs from various sources (see e.g. ref.8) has revealed that there exist distinct biases in the use of synonymous codons, and several arguments were noted concerning possible origins of the non-random usage of synonymous codons. Fitch (9) has analysed the sequence of MS2 phage RNA and has concluded, on the basis of the statistically significant excess of C over U in the third nucleotide position of codons, that selection has occurred against wobble pairing in pyrimidine-restricted codons. Berger (10) has compared the patterns of synonymous codon usage in four species of organisms, and has noted that for many amino acids the choice among synonymous codons is not common to all species and appears to be random, but for other codons specifying the amino acids such as arginine, isoleucine and glycine, distinct biases do exist in codon usage, which can be explained in terms of tRNA availability. In a statistical analysis of four mRNAs, Jukes (11) has shown a predominance of C and G in the third position of codons, and has inferred that this disproportionality results from DNA replicative errors which favour G-C pairing over A-T pairing.

Recently, Grosjean *et al.* (12) found that, in MS2 RNA, wobble codons are positively selected in codons having C and/or G in the first two posi-

tions and are selected against in codons having U and/or A in the first two positions. They concluded that this bias in codon usage in MS2 phage can be explained partially by considerations of the stability of the codon-anticodon complex in prokaryotic systems. Finding the opposite situation in the lacI gene of E.coli, Fiers and Grosjean (13) claimed that the difference of selective pattern of codon usage between the two systems should be related to a difference in translational rate with a proper amount of codon-anticodon interaction energy.

None of the above mentioned explanations has yet been verified by experiment. We think that the pattern of degenerate codon usage should be determined by many factors. We propose here another explanation that, in MS2 phage genome, the observed biases in codon assignment correlate well with the stability of the secondary structure of the RNA, a possibility suggested by Fitch (14) using a different method of analysis.

THE ANALYSIS OF MS2 RNA

Fiers and his coworkers (1-3) have determined the complete nucleotide sequence of MS2 RNA and have proposed a plausible secondary structural model, so-called flower model, of the RNA. Their proposal for the secondary structure may be in error in some minor details (15), but must be largely correct on the basis of the consideration of: 1, thermodynamics; 2, nuclease cleavage sites; and 3, the nature of nucleotide substitutions in related phages (16). Concerning the last consideration, we examined more in detail: among MS2, R17 and f2 phages, all the observed substitutions in pairing sites are those of the type that preserve base pairing (i.e., G-C or A-U versus G-U pairings) with only three exceptions resulting in the destruction of base pairing, where two of them, in sites 2568 and 2606, are at an end and at a beginning, respectively, of long stems and do not exert large influences upon the secondary structure.

In order to investigate a correlation between the bias in codon usage and the secondary structure of MS2 RNA, each nucleotide site of the RNA is assigned to pairing site or non-pairing site on the basis of the secondary structure proposed by Fiers et al.. Table 1 shows the observed frequencies n_{α}^{PS} and n_{α}^{NPS} of the nucleotide in the third position of codon α , the nucleotide being in pairing sites and in non-pairing sites, respectively, of MS2 coding regions. In most of the synonymous codon groups, significant excess of C and/or G over U and/or A is seen in pairing sites, which is in striking contrast to the trend in non-pairing sites. Table 2 shows the distribution

Table 1. The codon usage of MS2 phage, the third codon positions being pairing or non-pairing sites of the RNA.

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

PS: Pairing site, NPS: Non-pairing site.

Table 2. The distribution of nucleotides in the third codon positions of MS2 RNA, the positions being pairing or non-pairing sites.

Site		U	A	C	G	Sum
Pairing	Observed	154	113	246	183	696
	Expected	187.4	140.6	218.6	149.4	
Non-pairing	Observed	122	94	76	37	329
	Expected	88.6	66.4	103.4	70.6	
	Sum	276	207	322	220	1025

$\chi^2 = 69.7, \text{d.f.} = 3, P < 10^{-13}$

of nucleotides in the third position of codons in MS2 RNA, the position being pairing or non-pairing sites, where nondegenerate codons for methionine and tryptophan are excluded. The observed distribution differs considerably from the distribution expected when the pattern of codon usage does not differ

between pairing and non-pairing sites. There is a significant bias in the third codon position of MS2 RNA toward the use of C and G over U and A nucleotides in pairing as compared to non-pairing sites.

In order to show the above mentioned trend more clearly for individual groups of synonymous codons, it will be more appropriate to introduce an index representing a frequency of actual use relative to the frequency expected from random use of synonymous codons. We define an index "degree of preferable codon usage" P_{α}^{PS} (P_{α}^{NPS}) for a codon α whose third position is in a pairing site (non-pairing site) as follows:

$$P_{\alpha}^{PS} = n_{\alpha}^{PS} / \left(\sum_{\alpha' \in A} n_{\alpha'}^{PS} / d \right) \quad (1)$$

where n_{α}^{PS} is the observed frequency of the nucleotide in the third position of codon α in pairing sites (see Table 1), and sum runs over all the synonymous codons of the group specifying amino acid A with the degree of degeneracy d. For P_{α}^{NPS} in non-pairing sites, it is needed to substitute n_{α}^{NPS} for n_{α}^{PS} in the above equation. Most degenerate codons differ from each other only in the third position of codons except six-fold degenerate codon groups, which we tentatively divide into two-fold and four-fold degenerate codon groups for simplicity; 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 indicates any one of the four nucleotides, and Y and R indicate pyrimidine and purine, respectively. According to this classification, selective biases in the use of synonymous codons can be analysed in terms of composition of nucleotides in the third position of codons. Fig. 1 shows a plot of P_{α}^{NPS} versus P_{α}^{PS} for all the codons except nondegenerate ones. Most points, except Ile(AUU), Pro(CCC), Val(GUA), Val(GUG) and Arg(CGA), are in the region $P_{\alpha}^{NPS} > P_{\alpha}^{PS}$ for codons ending in U or A, and $P_{\alpha}^{NPS} < P_{\alpha}^{PS}$ for C or G. Moreover, there is an apparent tendency that the majority of the codons ending in C or G are plotted in the region of $P_{\alpha}^{PS} > 1$, $P_{\alpha}^{NPS} < 1$, and the codons ending in U or A are in $P_{\alpha}^{PS} < 1$, $P_{\alpha}^{NPS} > 1$. The average values of P_{α}^{PS} for U, C, A and G are 0.829, 1.308, 0.695 and 1.160, respectively, which are quite a contrast to that the average values of P_{α}^{NPS} for U, C, A and G are 1.365, 0.853, 1.255 and 0.457, respectively. That is, there is a selective code word usage that prefers C and/or G to U and/or A in the third codon positions for pairing sites, and vice versa for non-pairing sites. This nature implies that there exists a distinct bias for synonymous codon usage which stabilises the secondary structure of MS2 RNA, because G-C pairing is thermodynamically more stable than A-U pairing.

Fig. 1. Plot of P^{NPS} versus P^{PS} for degenerate codons. P^{PS} and P^{NPS} represent degrees of preferable usage of codon α whose third position is in a pairing and in a non-pairing site, respectively. Δ : a codon ending in U, $\blacktriangle$: C, $\circ$: A, and $\bullet$: G. The solid line shows a case $P^{NPS} = P^{PS}$, and the interrupted lines represent $P^{NPS} = 1$ and $P^{PS} = 1$, respectively.

Discussion

The present analysis based on the secondary structure of MS2 RNA shows that, in pairing sites of the RNA, there is a bias in the use of synonymous codons that favours C and/or G over U and/or A in the third codon positions, which is in striking contrast to the bias observed in non-pairing sites that favours U and/or A over C and/or G. On the basis of the predominance of C over U in the third position of pyrimidine-restricted codons of MS2 phage, Fitch (9) has suggested that selection has operated against wobble pairing between a codon and an anticodon of a tRNA. If his argument is correct, this disproportionality must be observed at all the third positions of codons, irrespective of the nucleotides being in base pairing sites or not, but it is not the case. A comparative analysis of nucleotide substitutions assures that selection has acted to favour the substitutions that would be compatible with the retention of the secondary structure. This implies that

this structure is not a temporal one, but that, during the evolution of the phage, selection has operated to conserve the existing structure. Two alternative explanations of the origin of the bias are possible; *i.e.*, one is that the present day secondary structure of MS2 RNA was the best among other alternative structures and the evolution has proceeded so as to stabilise this structure through biasing the codon usage, and the other is that the bias of the codon usage existed at first for some different unknown reasons, the secondary structure was then formed according to the bias so as to make more G-C pairings in the given sequence, and the resulting secondary structure survived during evolution. We think that the former interpretation is more plausible, but the latter possibility cannot be excluded. We have no clue now to settle the issue. We can only say that the bias in codon usage in MS2 genome is closely related to the stability of the secondary structure of the RNA.

The argument mentioned in this paper might be peculiar to single-stranded RNA phages such as MS2, R17 and f2, in which mRNA itself carries genetic information and, therefore, the single-stranded RNA should have a stable spatial structure. However, it is likely to be applicable also to many other mRNAs such as globin (17, 18) and histone H4 (19) mRNAs in which extensive secondary structures are expected to exist. Recently, Heindell *et al.* (20) have sequenced rabbit α -globin mRNA, and have proposed its secondary structure, although not yet established by experiments. A tentative analysis of their data is indeed in favour of our model; *i.e.*, in the third positions of degenerate codons, U or A are used 5 times, and C or G 67 times in pairing sites, and U or A 14 times, C or G 53 times in non-pairing sites, and the nucleotide usage in the third codon positions is significantly different between pairing and non-pairing sites ($\chi^2 = 5.72$, d.f. = 1, $P < 0.02$).

ACKNOWLEDGEMENTS

We thank Professor H. Matsuda for encouragement. We also thank a reviewer of this journal for suggesting the presentation of Table 2 and for valuable comments. This work was supported in part by a grant from the Japanese Ministry of Education, Science and Culture.

[†]To whom reprint requests should be sent.

REFERENCES

1. Min Jou, W., Haegeman, G., Ysebaert, M. and Fiers, W. (1972) *Nature* 237, 82-88.
2. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Merregaert, J., Min Jou, W., Raeymakers, A., Volckaert, G., Ysebaert, M., Van de Kerckhove, J., Nolf, F. and Van Montagu, M. (1975) *Nature* 256, 273-278.
3. 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. and Ysebaert, M. (1976) *Nature* 260, 500-507.
4. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. and Smith, M. (1977) *Nature* 265, 687-695.
5. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) *Science* 200, 494-502.
6. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) *Nature* 273, 113-120.
7. Beck, E., Sommer, R., Auerswald, E.A., Kurz, Ch., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M. (1978) *Nucleic Acids Res.* 5, 4495-4503.
8. Grantham, R. (1978) *FEBS Letters* 95, 1-11.
9. Fitch, W.M. (1976) *Science* 194, 1173-1174.
10. Berger, E.M. (1978) *J. Mol. Evol.* 10, 319-323.
11. Jukes, T.H. (1978) *J. Mol. Evol.* 11, 121-127.
12. Grosjean, H., Sankoff, D., Min Jou, W., Fiers, W. and Cedergren, R.J. (1978) *J. Mol. Evol.* 12, 113-119.
13. Fiers, W. and Grosjean, H. (1979) *Nature* 277, 328.
14. Fitch, W.M. (1974) *J. Mol. Evol.* 3, 279-291.
15. Berzin, V., Borisova, G.P., Cielens, I., Gribanov, V.A., Jansone, I., Rosenthal, G. and Gren, E.J. (1978) *J. Mol. Biol.* 119, 101-131.
16. Min Jou, W. and Fiers, W. (1976) *J. Mol. Biol.* 106, 1047-1060.
17. Favre, A., Morel, C. and Scherrer, K. (1975) *Eur. J. Biochem.* 57, 147-157.
18. Holder, J.W. and Lingrel, J.B. (1975) *Biochemistry* 14, 4209-4215.
19. Grunstein, M., Schedl, P. and Kedes, L. (1976) *J. Mol. Biol.* 104, 351-369.
20. Heindell, H., Liu, A., Paddock, G.V., Studnicka, G.M. and Salser, W.A. (1978) *Cell* 15, 43-54.