J. Mol. Evol. 12, 113-119 (1978)

Journal of Molecular Evolution

Bacteriophage MS2 RNA: A Correlation Between the Stability of the Codon: Anticodon Interaction and the Choice of Code Words*

Henri Grosjean[1], David Sankoff[2], Willy Min Jou[3], Walter Fiers[3], and R.J. Cedergren[2]

[1] Laboratoire de Chimie Biologique, Université de Bruxelles, 67, rue des Chevaux, B - 1640 Rhode St.-Genèse, Belgium

[2] Centre de Recherches Mathématiques et Département de Biochimie, Université de Montréal, Case postale 6128, Montréal, Québec, Canada

[3] Laboratory of Molecular Biology, University of Ghent, 35, Ledeganckstraat, B - 9000 Ghent, Belgium

Summary. The non-random distribution of degenerate code words in Bacteriophage MS2 RNA can be explained partially by considerations of the stability of the codon-anticodon complex in prokaryotic systems. Supporting this hypothesis we note that wobble codons are positively selected in codons having G and/or C in the first two positions. In contrast, wobble codons are statitically less likely in codons composed of A and U in the first two positions. Analyses of nucleotides adjacent to 5' and 3' ends of codons indicate a nonrandom distribution as well. It is thus likely that some elements of RNA evolution are independent of the structural needs of the RNA itself and of the translated protein product.

Key words: MS2 RNA – Codon-anticodon – Choice of code words

Introduction

The availability of the complete nucleotide sequence of Bacteriophage MS2 RNA has permitted the direct comparison of genetic information and corresponding gene products (Fiers et al., 1976). One of the observations made possible by analysis of this sequence is that the choice of degenerate codons is clearly non-random (Fiers et al., 1971 and 1976).

Since the first segments of the RNA molecule which were sequenced involved strongly basepaired hairpins, as these could be most readily obtained in pure form, it was originally thought that denegerate codons may be selected in order to optimise the secondary and tertiary structural stability of the RNA (Adams et al., 1969; MinJou et al., 1971). Further sequence information, however, did not provide statistically significant support for this hypothesis. Nevertheless, the secondary structure of MS2 RNA is decidedly more extensive and stronger than that of a random polynucleotide (7-9). As the mutation rate is higher in unpaired regions, (MinJou and Fiers, 1976), it follows that the secondary (and tertiary) structure must also be biologically important. It was then

* This work was supported by grants from the Belgian Government "Actions concertées - Gekoncertéerde Acties", N.F.W.O. and F.K.F.O. as well as from le Ministère de l'éducation du Québec. A preliminary report of this work was given at the EMBO ribosome workshop, Brussels 1976

0022-2844/78/0012/0113/$01.40

proposed by Ball (1973) that the higher than random secondary structure in phage RNA is brought about by constraints on the amino acid sequence, but this hypothesis has been criticized as well (Fitch, 1974). The only alternative is that the choice between degenerate codons for a given amino acid is in part influenced by structural requirements of the RNA. Such a forced choice, although sufficient to lead to the required secondary structure, would not necessarily be detectable by statistical analysis. (Fiers et al., 1976).

Another possible constraint on the free choice of degenerate code words is that of modulation control, i.e., that the translation rate of different cistrons is regulated in addition to the initiation frequency) by the use of certain rate-limiting codons (Ames and Hartman, 1963; Stent, 1964). These would be recognized either by a minor species of an isoaccepting tRNA or would interact inefficiently with their cognate tRNA. Indeed, in MS2 RNA there is a remarkable correlation between the high efficiency of coat protein translation and the absence of certain code words, which are precisely recognized by a minor species of isoaccepting tRNA. Particularly the codons AUA for isoleucine and AGA/AGG for arginine are candidates for such a modulation role. There is also some genetic evidence which pleads for a modulation role of AUA (MinJou et al. 1976).

But for many amino acids, the choice between the alternative codons is clearly nonrandom, as revealed by χ-square analysis (Fiers, 1976), and yet the constraints cannot be adequately explained by the effects discussed above. Therefore, one can speculate that another important factor which leads to nonrandom codon use, may be dictated by the requirements of the translation machinery and, in particular the efficiency of codon-anticodon interaction. Indeed, several observations point to the existence of well-defined structural rules in this interaction. From data on tRNA sequences (Sprinzl et al., 1978) it is obvious that all possible combinations of the four bases A, G, C, U are not found in the 3 positions of the anticodon. For example, A is never present in the wobble position of the tRNA anticodon and consequently all codons ending with a pyrimidine base must be read by an anticodon starting with a guanosine or inosine. Furthermore, in *E. coli*, and most probably in all prokaryotic cells, no tRNAs, which have a G in the second position of the anticodon, contain C in the first position. In addition U found in the first (wobble) position is almost always modified as in uridine-5-oxyacetic acid or derivatives of 2-thiouridines (McCloskey and Nishimura, 1977). In the latter cases the thiolation of uridine considerably increases its stacking energy (Mazumdar et al., 1974) rendering the modified nucleoside incapable of wobbling, i.e., reading a G in third position of the codon (Ohashi et al., 1970 et Grosjean et al., 1978). More striking is the relationship between the nucleoside in the third position of the anticodon and 3'-adjacent nucleoside. In *E. coli* tRNA, when the third nucleoside is A or U the adjacent nucleoside is invariably hypermodified as in 2-methylthio-6-isopentenyl adenosine or 6-threonyl adenosine except for initiator tRNA. Recent studies on the stability of anticodon-anticodon interactions, show these hypermodified nucleosides contribute significantly to the thermal stability of a base-paired complex, especially those involving the weaker A-U pair (Grosjean et al, 1976). In fact, the proper functioning of these tRNAs in prokaryotic ribosome-directed protein synthesis is contingent on the presence of the hypermodified nucleoside adjacent to the anticodon, although the aminoacylation reaction is not (Gefter and Russell,

1969; Miller et al., 1976). *E. coli* tRNAs, which contain only G and/or C in the anti-codon, (glycine, alanine, and proline) have no such hypermodification suggesting that the intrinsic stability of their codon-anticodon complexes is sufficient for proper translation.

These observations lead one to the notion of an "optimal" energy of stabilisation for the codon-anticodon complex. In the present study we show that this optimal energy concept can explain at least part of the non-random distribution of codon frequencies in bacteriophage MS2 RNA.

The Analysis of MS2 RNA

In a first series of analyses we have compared triplet frequencies in all three genes of MS2 RNA (1068 codons assigned) with expected values calculated from the base analysis of each triplet position. In this way both in phase triplets (the codon) denoted by 1, 2, 3 and out of phase triplets consisting either of the last two positions in one codon and the first position in the next, denoted by 2, 3.1 or the last position plus the first two in the next codon, denoted by 3. 1,2- have been tabulated.

A chi-squared test of how non-random are the data in these three contingency tables gives: 1, 2, 3.; $\chi^2 = 166.8$; for 2, 3.1 $\chi^2 = 69.4$ and for 3.1, 2 $\chi^2 = 37.1$. It is thus obvious that an important contribution to the skewed distribution in the MS2 sequence involves the non-random in-phase codon utilization, a large but not predominant proportion of which results from the absence of the termination codons, UGA, UAA, and UAG.

Further analysis of code word use is complicated by the constraints imposed on the nucleotide sequence by the aminoacid requirements of the resulting protein. For example, the overuse of the tryptophan codon UGG (23 times vs 11 times expected for a random distribution) is probably due to the structural requirements of the proteins coded for by MS2 RNA.

Therefore, in the present analysis we have concentrated our effort on degenerate codons read by the same tRNA. In this way, we avoid restrictions imposed on the protein structure, since both codons are for the same aminoacid. Likewise we avoid any possible translation control imposed by cellular tRNA levels (discussed above) since the two codons are read by the same tRNA.

A list of codons, their frequencies and the anti-codon sequence of the *E. coli* tRNA which reads them is given in Table 1. We may then ask the question: is the use of the "wobble" interaction G-U instead of G-C dependent on the first two nucleotides in the codon? The answer is clearly yes. In codons containing A and/or U, in the first and second position, the wobble U base is strongly selected against. For example, the tyrosine codon UAU is used 9 times, whereas the UAC codon occurs 32 times. The codons of phenylalanine (UUU and UUC), isoleucine (AUU and AUC) and asparagine (AAU and AAC) show a similar preferential use of the NNC codon.

But in the case of codons containing G and/or C in the first and second position, the phenomenon is reversed. Here the codons containing the wobble U base are selected. Thus for glycine, the codon GGU is used 37 times and GGC only 16 times. The differences of equivalent codons for proline (CCU over CCC) and alanine (GCU over GCC) are less pronounced, but still evident. The other G-C containing codon (arginine) is

Table 1. The frequency of each codon used in MS2-RNA together with the anticodon sequence of the tRNA reading the particular codon. Codon frequencies come from complete analysis of MS2-RNA sequence (Fiers et al., 1976). Anticodon identification come from tRNA sequence analysis except for the anticodons GGA, VGA, and GGC which were identified from studies by anticodon-anticodon interaction (H. Grosjean et al., 1978). All other anticodons, the existence of which is expected from the properties of the *E. coli* translation system but that have not yet been identified are in brakets (8 of 42 anticodons). The four anticodons starting with G and having only A and/or U in the two first position of the anticodon are underlined. The three anticodons starting with G and having only G and/or C in the two first position of the anticodon are double underlined. The table does not include the three initiator codons (respectively two AUG and one GUG) read by F-MET-tRNA (anticodon CAU). V is uridin-5-oxyacetic acid, C^+ is N^4-acetyl cytidine, S is 5-methylaminomethyl-2-thiouridine, and Q is 7-(4,5-cisdihydroxy-1-cyclopenten-3-ylaminomethyl)-7-deazaguanosine, N is a unidentified modified nucleoside and U* is a unidentified modified uridine

	U				C				A				G			
U	PHE	19	GAA	SER	15	GGA	TYR	9	QUA	CYS	6	GCA	U
		29			20			32			6		C
	LEU	17	NAA		16	VGA	OCHRE	1	U*UA	OPAL	0		A
		11	NAA		22		AMBER	2	CUA	TRP	23	CCA	G
C	LEU	15	GAG	PRO	17	(GGG)	HIS	6	QUG	ARG	21	ICG	U
		26			10			9			20		C
		15	(NAG)		9	VGG	GLN	17	SUG		10		A
		9	CAG		13			22	CUG		11	(CCG)	G
A	ILE	12	GAU	THR	19	GGU	ASN	17	QUU	SER	8	GCU	U
		25			21			28			16		C
		19	U*AU		13	(VGU)	LYS	19	SUU	ARG	7	(NCU)	A
	mMET	18	C^+AU		14			26	(CUU)		6	(CCU)	G
G	VAL	21	GAC	ALA	26	GGC	ASP	28	QUC	GLY	37	GCC	U
		21			21			22			16		C
		16	VAC		21	VGC	GLU	16	SUC		12	U*CC	A
		18			23			28	(CUC)		16	CCC	G

particular, at least in *E. coli*, in that the $tRNA^{Arg}$ contains an inosine instead of a G in the wobble position.

It seems, therefore, that in the case of intrinsically weak codon-anticodon interaction, involving A and/or U in position 1 and 2, it is the strong, non-wobble codon which is strongly favored. The opposite effect is found in intrinsically strong codon-anticodon interactions, that is containing G and/or C in the first two positions, here the weak wobble interaction is favored. The interpretations that we give here to explain codon choice are consistent and compatible with the notion of the *optimal* energy requirement of codon-anticodon interaction previously mentioned. Although Table 1 was constructed from the sum of codons from the three MS2 proteins, the correlation between codon composition and use of the wobble holds for three MS2 genes individually.

The correlation does not hold for codon pairs containing purines in the third position of the codon which are read by the same tRNA having a V base at the wobble position in anticodon. All these tRNAs (specific for serine, proline, threonine, alanine and valine whose contain at least one G or C in the first two positions of anticodon might already fall within the range of "optimal" interaction with the corresponding codon. For these codons indeed no preference is evident (see Table 1).

It may be significant, however, that intrinsically weak codons, i.e., composed of A and/or U in the first two positions have a strong preference for Watson-Crick pair in the third position.

Subsequently we examined the skewness of out-of-phase triplet frequencies (i.e., 3.1, 2 and 2, 3.1) for any interpretable tendencies. While there is clearly a non-random use of nucleotides adjacent to the 3'-end of codons as indicated by the χ square value cited above we have not been able to interpret this in any coherent manner. For example, we take all 16 codons ending in G and divide them in two groups according to whether they are read by the same tRNA as the corresponding codon termination in A (like with codons specific for valine, serine, proline, threonine and alanine), or whether they are read by different tRNAs. In the first case, a typical non-random frequency of the 3' adjacent base is found: A, 32 times; C, 22; G, 19; U, 27. While in the second case we found A, 40; C, 39; G, 55; U, 38. The ratio of G in the two cases is striking but as yet uninterpretable. The non-random distribution of out-of-phase triplets could be significant especially in connection with recent work on the variation of suppressor tRNA efficiencies as a function of "context" of the suppressible chain terminating codons (Akaboschi et al., 1976; Colby et al., 1976; Feinstein and Altman, 1977). If context in mRNA is important to the proper codon-anticodon interation one would expect a relationship between codons and their preceding and following bases. Context effects are of course already shown on the tRNA side of the codon-anticodon interaction (see above on the hypermodified nucleotide).

In conclusion, the finding that degenerate codon selection may depend on the energetic nature of the codon-anticodon complex has some important implications in the study of the translational process.

1. The mRNA sequence evolves at least to some extent independently of restrictions imposed by the protein products activity or structure and by the secondary and tertiary structural requirements of the RNA (see Fitch, 1976). It may be predicted,

however, that the wobble correlation from a prokaryotic system presented here could differ significantly from eukaryotes. One indication of this difference is the use of more inosine in the wobble position of eukaryotic tRNAs.

2. The biological rationale of the wobble in MS2 RNA decoding may be related to a "fine tuning" of the energy requirement of a proper codon-anticodon interaction. This hypothesis should be testable using thermodynamic studies such as those already in progress by one of us (H.G.)

3. The finding that nucleotides adjacent to codons are not randomly distributed would suggest that "context" effects may play an important role in proper translation. This suggestion is lent credibility by reports on the context effect on suppression of nonsense mutations.

Finally, one important question is to know if our observation made with Bacteriophase MS2 RNA can be generalized to other prokaryotic mRNA. Unfortunately, at present it is not possible to compare the results with enough long sequences of prokaryotic mRNAs. However, a comparison can be made with the nucleotide sequence of the phage ΦX174 (1346 codons unambiguously assigned, Sanger et al., 1977). The use of synonymous codons is clearly non-random, but the pattern is quite different from that observed with MS2 RNA. There is an extremely striking preference for U in the third position. This may be due to other constraints, which are more important than an optimalisation of the translation efficiency, for example, synthesis of single standed DNA or packaging. Also, the burst size of a ΦX174 infection is only about 200 - 400 particles per cell (Sinsheimer, 1970) while it reaches 5,000 to 10,000 for the RNA-phages, again pointing to the remarkable translation (and replication) efficiency of the latter. We do note, however, that also in the case of ΦX174 the codons, which we have previously implicated in a modulation function, are conspicuously low, viz. AUA for isoleucine and AGA/AGG for arginine. This comparison does not mean that the phenomenon of an optimal codon-anticodon interaction is restricted to MS2 RNA. Indeed, the doublet frequencies, especially of the MS2 coat gene, are remarkably similar to those of the host cell genome, Elton et al., 1976).

References

Adams, J.M., Jeppesen, P.G.N., Sanger, F., Barrell, B.G. (1969). Nature **223**, 1009

Akaboschi, E., Inoye, M., Tsugita, A. (1976). Mol. Gen. Genet. **149**, 1

Ames, B.N., Hartman, P.E. (1963). Cold Spring Harbor Symposia **28**, 349

Ball, L.A. (1973). Nature New Biol. **242**, 44

Colby, D.S., Schedl, P., Guthrie, C. (1976). Cell **9**, 449

Elton, R.A., Russell, G.J., Subak-Sharpe, J.H. (1976). J. Mol. Evol. **8**, 117

Feinstein, S.I., Altman, S. (1977). J. Mol. Biol. **112**, 453

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., MinJou, W., Molemans, F., Raeymakers, A., Berghe, A., Van den, Volckaert, G., Ysebaert, M. (1976). Nature **260**, 500

Fiers, W., Contreras, R., Wachter, R., De, Haegeman, G., Merregaert, J., MinJou, W., Berghe, A., Van den (1971). Biochimie **53**, 495-506

Fitch, W.M. (1974). J. Mol. Evol. **3**, 279

Fitch, W.M. (1976). Science **194**, 1173

Gefter, M.R., Russell, R.L. (1969). J. Mol. Biol. **39**, 145

Grosjean, H., Henau, S., de, Crothers, D. (1978). Proc. Natl. Acad. Sci. **75**, 610
Grosjean, H., Söll, D., Crothers, D. (1976). J. Mol. Biol. **103**, 499
Mazumdar. S.K., Saenger, W., Scheit, K.H. (1974). J. Mol. Biol. **85**, 213
McCloskey, J.A., Nishimura, S. (1977). Accts. Chem. Res. **10**, 403
Miller, J.P., Hussain, Z., Schweizer, M.P. (1976). Nucleic Acid Res. **3**, 1185
MinJou, W., Fiers, W. (1976). J. Mol. Biol. **106**, 1047
MinJou, W., Haegeman, G., Fiers, W. (1971). FEBS Letters **13**, 105
MinJou, W., Montagu, M., Van, Fiers, W. (1976). Biochem. Biophys. Res. Comm. **73**, 1083
Sanger, F., Air, G.N., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchinson, C.A., Slocombe, P.M., Smith, M. (1977). Nature **265**, 687
Sinsheimer, R.L. (1970). In: The Harvey Lectures. New York: Academic Press, series 64, p. 69
Sprinzl, M., Grüter, F., Gauss, D.H. (1978). Nucl. Acid. Res. **5**, r 15
Stent, G.S. (1964). Science **144**, 816

Received May 19, 1978/Revised September 11, 1978