(Reprinted from Nature, Vol. 260, No. 5551, pp. 500–507, April 8, 1976)

Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene

W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert & M. Ysebaert

Laboratory of Molecular Biology, University of Ghent, 9000 Ghent, Belgium

Bacteriophage MS2 RNA is 3,569 nucleotides long. The nucleotide sequence has been established for the third and last gene, which codes for the replicase protein. A secondary structure model has also been proposed. Biological properties, such as ribosome binding and codon interactions can now be discussed on a molecular basis. As the sequences for the other regions of this RNA have been published already, the complete, primary chemical structure of a viral genome has now been established.

RNA BACTERIOPHAGES such as MS2 have been important in molecular biology not only because they provide a model system for investigating viral RNA replication and the physiology of the infected cell, but also in the study of fundamental processes such as translation[1]. MS2 RNA contains the genetic information to specify three viral polypeptides (Fig. 1). The first two genes code for the A protein and the coat protein, which are structural elements: the virus particle contains one A-protein molecule[2] and approximately 180 coat protein molecules. For practical reasons, most knowledge of the *in vitro* replication of phage RNA is based on work with a purified system derived from cells infected with the distantly related bacteriophage Qβ (reviewed in ref. 3), but the principal findings are presumably also valid for the MS2 phage group[4]. The enzyme complex responsible for viral RNA replication has been referred to variously as viral RNA-dependent RNA polymerase[5], RNA synthetase[6] and replicase[5]. As the latter name is simple and unambiguous we use it here. The replicase complex consists of four polypeptides, α, β, γ and δ[4,7,8]. β is specified by the third and last viral gene and has a molecular weight of approximately 63,000 (ref. 4). α has been identified as the 30S ribosomal subunit S1 (ref. 9) and γ–δ as the elongation factors Tu–Ts[1,3].

The expression of the three viral genes is strictly regulated. The A-protein gene is presumably only translated from chains in a nascent state[1,10,11]. The replicase gene is subject to a polar control by the coat gene; this means that the latter has to be translated to allow expression of the replicase gene[1,10]. Moreover, the coat protein represses the replicase gene[1,10] while the replicase complex in turn might be involved in shutting off coat gene expression[12]. Regulation of the frequency of gene expression by modulation has also been proposed[13,14].

We have shown that the MS2 genome starts from the 5′ end with a 129-nucleotide untranslated leader sequence (Fig. 1)[15]. Then follows the A-protein gene which starts with a G–U–G initiation codon and ends with a U–A–G codon[11,16], an intercistronic region of 26 nucleotides, the coat protein gene[14], an intercistronic region of 36 nucleotides[14,17], the replicase gene (see below) and finally a 174-nucleotide untranslated segment at the 3′ terminus[16,18]. The proposed secondary structure models provide a rational basis for explaining such biologically relevant phenomena as the autocontrol of A-protein gene expression[11], the polarity effect[14] and the location of easily mutable sites[11,14].

We now report the complete primary structure of the third and last MS2 gene, which codes for the replicase subunit. Some partial nucleotide sequences have been published before[13,19,20].

Nucleotide sequence and structure

Our methods, which have been described before[11,16,18], were briefly as follows. Uniformly ^{32}P-labelled MS2 RNA (2–8 mCi) was digested partially at 0 °C with the single-strand-specific ribonuclease T_1, and the digest was fractionated by electrophoresis on a polyacrylamide slab gel[15]. A series of bands was obtained, and the material in each band was separated further into individual fragments by two-dimensional gel electrophoresis[21]. The advantages of the latter method are that, in general, pure and unique fragments are obtained and that the procedure is applicable to virtually any chain length. These pure fragments were suitable for detailed sequence analysis, mainly according to the methodology described by Sanger and colleagues[22,23], or modifications thereof[15,24]. The structural determination of the longer oligonucleotides required various methods[25,26] and will be described in detail elsewhere.

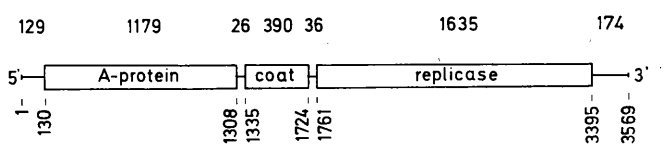


Fig. 1 The genome of MS2 contains three genes, shown in blocks. Untranslated regions are present at both ends of the viral RNA and between the genes. The length of the different regions, expressed in number of nucleotides, is shown on top. The nucleotides in the viral RNA are numbered from the 5′ end to the 3′ end and the positions of some important signals are indicated underneath (the initiation codon is considered part of the genes and the termination codon part of the untranslated region).

Discrete fragments can be obtained reproducibly because of the specific three-dimensional conformation of the viral RNA; for example, a fragment may correspond to a hairpin. In milder conditions of partial enzymatic digestion, larger fragments are obtained, for example, several hairpins linked to each other or bound by secondary interactions. Although in this way several regions can often be ordered, this approach has practical limits. On the one hand, if a fragment becomes too large (for example, more than 300–400 nucleotides) its oligonucleotide composition can no longer be established accurately and on the other hand, some sites in the viral RNA are extremely sensitive to T_1 ribonuclease (see legend to Fig. 2). To solve this problem, we introduced the use of carboxymethylated ribonuclease (CM ribonuclease); this enzyme cleaves mainly between C and A

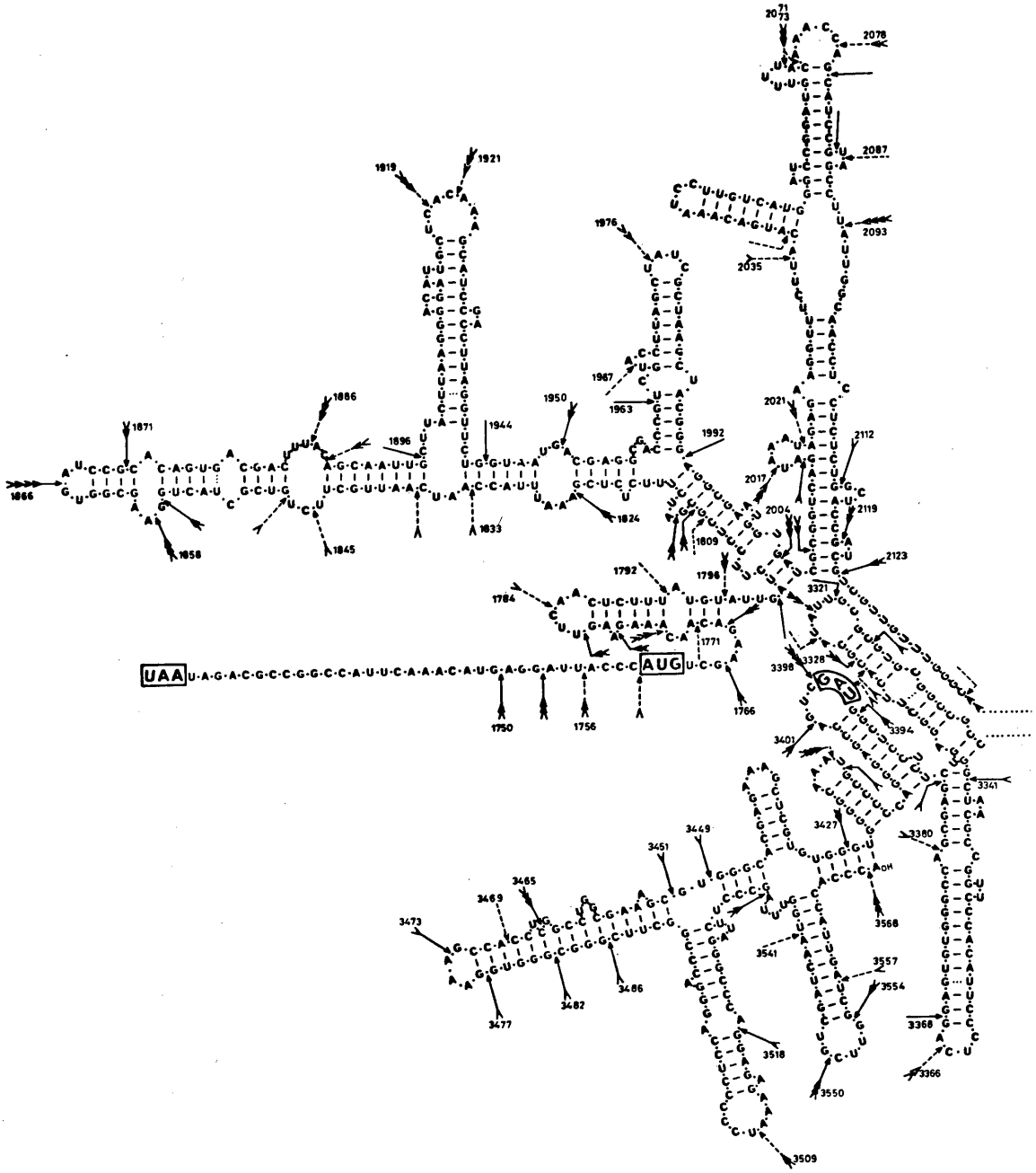
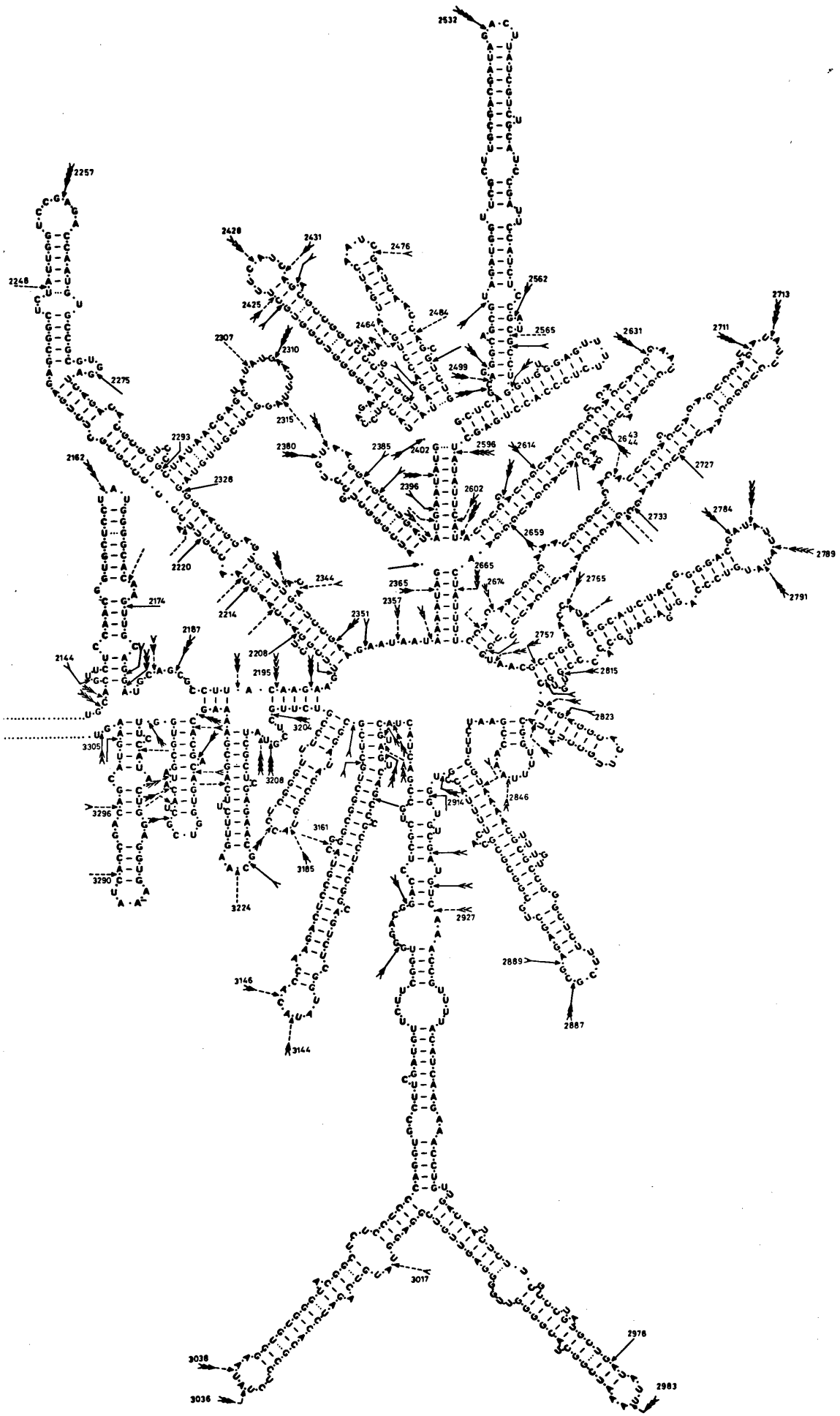


Fig. 2 Primary nucleotide sequence and secondary structure model of the replicase gene. The sequence is read from the 5′ to the 3′ end: U–A–A, termination signal of the preceding coat gene; an intergenic region of 33 nucleotides; A–U–G, the initiation codon of the replicase gene; the replicase gene (1,632 nucleotides); U–A–G, the termination signal; an untranslated, 171-nucleotide 3′-terminal sequence. Together with previous models for the 5′-leader sequence–A-protein gene[11] and the coat gene[14] this figure completes the chemical structure of the entire MS2 RNA. The numbering starts at the first (5′) nucleotide and ends at the 3′-terminal residue 3569. The primary structure has been established rigorously but the secondary structure is based in part on experimental evidence and theoretical predictions and in part on conjectures (see text). Arrows point to sites easily split by nucleases during partial digestion of complete MS2 RNA: solid arrows, T_1 ribonuclease; dashed arrows, CM ribonuclease. The feathers indicate the sensitivity of the bond in the conditions used: (0) split very seldom, (1) seldom, (2) rather often, (3) very often, (4) always.


2532
2257
2428
2431
2476
2562
2248
2425
2484
2464
2565
2631
2713
2711
2307
2310
2499
2275
2315
2293
2380
2385
2402
2596
2614
2643
2644
2727
2162
2326
2396
2602
2733
2784
2659
2789
2220
2665
2674
2765
2791
2344
2365
2214
2357
2351
2174
2187
2208
2757
2144
2195
2815
2823
3305
3204
3208
2914
2846
3296
3161
3185
2927
3224
2889
3146
2887
3144
3017
2978
3038
2983
3036

and between U and A residues[27]. Such partial digestion of MS2 RNA with CM ribonuclease identified a complete new set of fragments. These not only often constituted suitable material for solving nucleotide regions not well represented in T_1 fragments, but by and large provided all the essential overlaps, which have enabled us to reconstruct the entire sequence.

For identification of the fragments and for correlating the T_1 fragments with the CM ribonuclease fragments, we used a catalogue of all the unique oligonucleotides (mainly T_1 oligonucleotides). In this respect, an improved minifingerprinting system[28] has been very helpful in the last stages of this work.

The nucleotide sequence of the replicase gene is shown in the form of a secondary structure model in Fig. 2. It starts with the initiation codon A–U–G and ends with the termination codon U–A–G; these regions have been published before[14,16,18]. Figure 2 also shows sites which are sensitive to ribonuclease T_1, but which could be bridged by suitable CM-ribonuclease fragments, and vice versa for sensitive CM sites. For example, the T_1 fragment 2381–2532 was nearly always found in the same band of the primary gel separation system as the T_1 fragment 2533–2631, but no linkage in primary sequence could be demonstrated. Direct proof for the correct order, however, was provided by the characterisation of several CM fragments, such as 2500–2596.

Such fragments, clearly linked by secondary interactions and only separated in the denaturing conditions of the first dimension of the two-dimensional gel fractionation system[21], provided the primary basis for the construction of the secondary structure model. Sometimes a unique pairing scheme could be derived, but more often there were several alternatives (usually varying only in detail). On the basis of simplified rules for estimating thermodynamic stability[29,30], the most plausible configuration can be selected. After these hairpins have been fixed, additional, thermodynamically stable interactions can then be sought. The model shown in Fig. 2 is much more tentative as far as the latter aspects are concerned. But it is satisfying that by and large the feathered arrows, which are a measure of the sensitivity of the site towards nucleases, point to loops at the top of hairpins, or to discontinuities in the helical segments or to other single-stranded loci. Further study of the model by chemical modification is in progress.

Some of the segments, which remained single stranded in the model shown in Fig. 2, can be base paired to other regions of the molecule (Fig. 3). Such long distance interactions (that is, linkage of segments far apart in the primary sequence) certainly agree with hydrodynamic and other physical properties of MS2 RNA[31,32], but they are not supported by direct experimental evidence. Some of these interactions may have important functional roles, as in the control of the genetic expression or in encapsidation. For example, the coat protein gene must be translated before the ribosome-binding region of the replicase gene becomes accessible. This polar effect can be explained by the interaction of segment 1409–1433 with segment 1738–1769 (Fig. 3), as proposed in a slightly different form[14]. This does not exclude that rearrangements to other conformations may occur[14,33].

Little can be said about the tertiary interactions of the type found in tRNA[34], which almost certainly are also present in MS2 RNA. Yet they may influence profoundly the structure–function relationship. Moreover, some alternative conformations which have not been retained in the present model (Fig. 2) may be stabilised by such tertiary interactions and thus may represent the true conformation.

Sequence of replicase subunit

On the basis of the nucleotide sequence, we can deduce the whole amino acid sequence of the replicase gene product (Fig. 4); only the first three amino acids had been established previously by direct peptide analysis[35,36]. This is the first protein for which the primary structure has been solved entirely on the basis of the genetic information which encodes it.

The replicase subunit is a 544-amino acid polypeptide. Like the A protein, it ends with an arginine residue. It is even more rich in arginine than the A protein (some arginine residues may be involved in interactions with the RNA). But as the aspartic acid content is also high (6.0%), the polypeptide is slightly less basic than the A protein. Also noteworthy are a relatively high leucine content (9.0%; also high in the A protein) and phenylalanine content (6.9%). On the other hand, valine, alanine and glutamine are low, compared with the two other viral proteins.

At present no very meaningful deductions can be made from the amino acid sequence. With available procedures it is not possible to derive a plausible conformation and even less to interpret the structure–function relationship. But when more

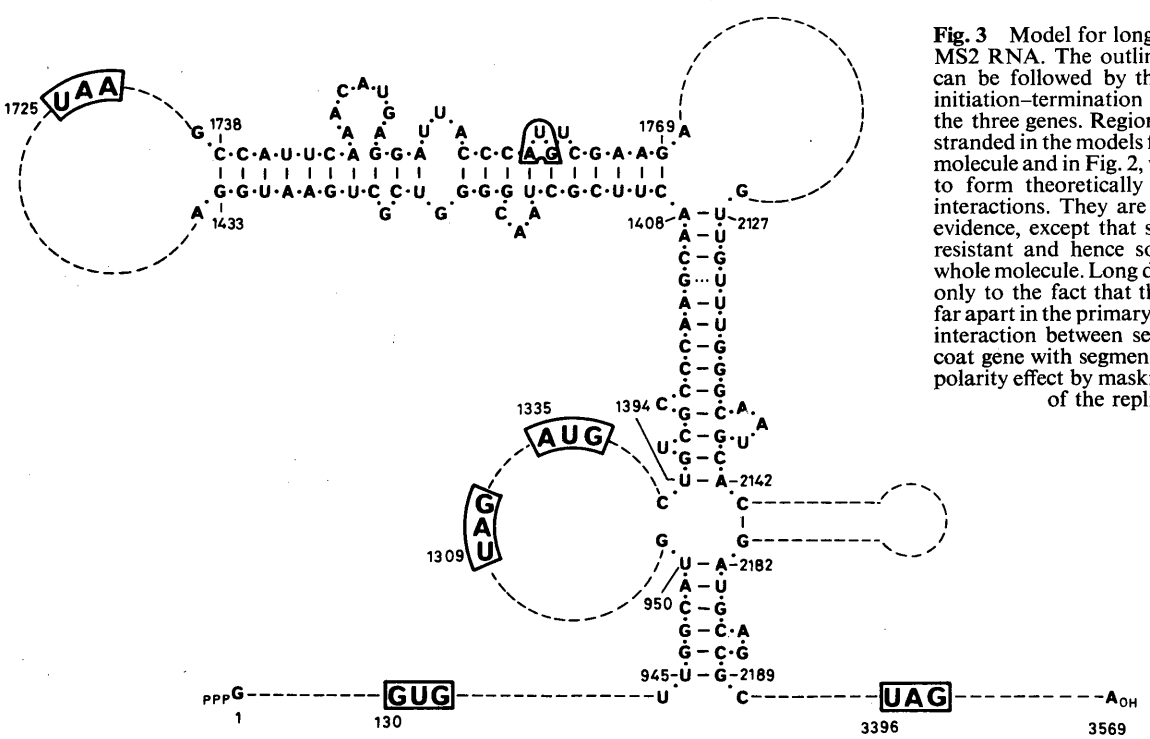


Fig. 3 Model for long distance interactions in MS2 RNA. The outline of the complete chain can be followed by the numbering and by the initiation–termination signals (in bold type) of the three genes. Regions which remained single stranded in the models for the other parts of the molecule and in Fig. 2, were screened for ability to form theoretically stable, complementary interactions. They are not supported by direct evidence, except that some are rather nuclease resistant and hence somehow protected in the whole molecule. Long distance interaction refers only to the fact that the segments involved are far apart in the primary sequence. Note that the interaction between segment 1409–1433 of the coat gene with segment 1738–1769 explains the polarity effect by masking the initiating A–U–G of the replicase gene.

data become available on viral and other RNA polymerases these amino acid sequences may reveal important clues on virus evolution and/or their origin.

Biological information content

The ribosome-binding region, including the initiating A–U–G of the replicase gene, was first identified by Argetsinger-Steitz[37]. Good yields were only obtained after partial unfolding of the viral RNA. It is still not clear which features constitute a ribosome-binding region, although progress has been made. Shine and Dalgarno[38] have proposed that a purine-rich sequence preceding the initiation codon is involved in an interaction with a complementary 3′-terminal segment of the ribosomal 16S RNA. Indeed, all genuine ribosome-binding sites characterised so far contain a sequence of three or more purine nucleotides complementary to the 3′ end of 16S RNA. Also, experimental evidence has been obtained for the direct complementary interaction between the ribosome-binding region and the 3′ end of the 16S rRNA[39]. But this cannot be the whole answer, because many regions in the viral RNA fulfil these criteria, yet are not bound. For example, the segment 2381–2396 is . . . U–A–A–G–G–A–G–C–C–U–G–A–U–A–U–G . . .; according to the criteria mentioned it should interact about as strongly as the A-protein initiation region with both *Escherichia coli* and *Bacillus stearothermophilus* 16S rRNA[38]. That this is not observed may mean that masking/accessibility by secondary and tertiary structure may be equally important. Such cryptic sites may be involved in the additional initiations observed after formaldehyde treatment[40], but these are very unnatural conditions. On the other hand, ribosome binding as tested by protection against nuclease may not be as specific as polypeptide chain initiation. In the original fingerprints of the protected regions of the *E. coli* ribosome there were additional, unexplained spots[37]; on the basis of preliminary identification (J. A. Steitz, personal communication) it seems likely that the main contaminating sequence contained the region 1965–1986. Likewise, in the experiment with *B. stearothermophilus* ribosomes at 49 °C (plate IV in ref. 41) a major contaminant must have been 2926–2967. In neither case is it immediately obvious why this anomalous binding reaction occurred.

Translation of the replicase gene is repressed by the coat protein, and the region to which the coat binds has been characterised[42]. The coat protein may interact with an alternative hairpin structure (Fig. 6 of ref. 14, and ref. 33).

Starting from the initiating A–U–G the nucleotides are read in triplets. For many regions the correct reading frame could be derived independently from the fact that the two illegitimate reading frames were blocked by nonsense code words[19]. The

Fig. 4 Nucleotide sequence of the replicase gene and amino acid sequence of the coded polypeptide. As in Figs 1–3, the nucleotides are numbered from the 5′ end of the viral RNA. The amino acid sequence is entirely deduced from the nucleotide sequence data. In analogy with the Qβ polypeptide it is assumed that the initiating formylmethionine has been eliminated in a post-translational process[55]. This polypeptide, together with three host components S1, Tu and Ts, forms a complex which is responsible for specific viral RNA replication.

MS2 REPLICASE GENE

```
1764
A-U-G U-C-G.A-A-G.A-C-A.A-C-A.A-A-G.A-A-G.U-U-C.A-A-C.U-C-U.U-U-A.U-G-U.A-U-U.G-A-U.C-U-U.C-C-U.C-G-C.G-A-U.C-U-U.U-C-U.C-U-C.
   Ser - Lys - Thr - Thr - Lys - Lys - Phe - Asn - Ser - Leu - Cys - Ile - Asp - Leu - Pro - Arg - Asp - Leu - Ser - Leu -
                                                                                                                        20
1824
   G-A-A.A-U-U.U-A-C.C-A-A.U-C-A.A-U-U.G-C-U.U-C-U.G-U-C.G-C-U.A-C-U.G-G-A.A-G-C.G-G-U.G-A-U.C-C-G.C-A-C.A-G-U.G-A-C.G-A-C.
   Glu - Ile - Tyr - Gln - Ser - Ile - Ala - Ser - Val - Ala - Thr - Gly - Ser - Gly - Asp - Pro - His - Ser - Asp - Asp -
                                                                                                                        40
1884
   U-U-U.A-C-A.G-C-A.A-U-U.G-C-U.U-A-C.U-U-A.A-G-G.G-A-C.G-A-A.U-U-G.C-U-C.A-C-A.A-A-G.C-A-U.C-C-G.A-C-C.U-U-A.G-G-U.U-C-U.
   Phe - Thr - Ala - Ile - Ala - Tyr - Leu - Arg - Asp - Glu - Leu - Leu - Thr - Lys - His - Pro - Thr - Leu - Gly - Ser -
                                                                                                                        60
1944
   G-G-U.A-A-U.G-A-C.G-A-G.G-C-G.A-C-C.C-G-U.C-G-U.A-C-C.U-U-A.G-C-U.A-U-C.G-C-U.A-A-G.C-U-A.C-G-G.G-A-G.G-C-G.A-A-U.G-G-U.
   Gly - Asn - Asp - Glu - Ala - Thr - Arg - Arg - Thr - Leu - Ala - Ile - Ala - Lys - Leu - Arg - Glu - Ala - Asn - Gly -
                                                                                                                        80
2004
   G-A-U.C-G-C.G-G-U.C-A-G.A-U-A.A-A-U.A-G-A.G-A-A.G-G-U.U-U-C.U-U-A.C-A-U.G-A-C.A-A-A.U-C-C.U-U-G.U-C-A.U-G-G.G-A-U.C-C-G.
   Asp - Arg - Gly - Gln - Ile - Asn - Arg - Glu - Gly - Phe - Leu - His - Asp - Lys - Ser - Leu - Ser - Trp - Asp - Pro -
                                                                                                                       100
2064
   G-A-U.G-U-U.U-U-A.C-A-A.A-C-C.A-G-C.A-U-C.C-G-U.A-G-C.C-U-U.A-U-U.G-G-C.A-A-C.C-U-C.C-U-C.U-C-U.G-G-C.U-A-C.C-G-A.U-C-G.
   Asp - Val - Leu - Gln - Thr - Ser - Ile - Arg - Ser - Leu - Ile - Gly - Asn - Leu - Leu - Ser - Gly - Tyr - Arg - Ser -
                                                                                                                       120
2124
   U-C-G.U-U-G.U-U-U.G-G-G.C-A-A.U-G-C.A-C-G.U-U-C.U-C-C.A-A-C.G-G-U.G-C-U.C-C-U.A-U-G.G-G-G.C-A-C.A-A-G.U-U-G.C-A-G.G-A-U.
   Ser - Leu - Phe - Gly - Gln - Cys - Thr - Phe - Ser - Asn - Gly - Ala - Pro - Met - Gly - His - Lys - Leu - Gln - Asp -
                                                                                                                       140
2184
   G-C-A.G-C-G.C-C-U.U-A-C.A-A-G.A-A-G.U-U-C.G-C-U.G-A-A.C-A-A.G-C-A.A-C-C.G-U-U.A-C-C.C-C-C.C-G-C.G-C-U.C-U-G.A-G-A.G-C-G.
   Ala - Ala - Pro - Tyr - Lys - Lys - Phe - Ala - Glu - Gln - Ala - Thr - Val - Thr - Pro - Arg - Ala - Leu - Arg - Ala -
                                                                                                                       160
2244
   G-C-U.C-U-A.U-U-G.G-U-C.C-G-A.G-A-C.C-A-A.U-G-U.G-C-C.G-C-G.U-G-G.A-U-C.A-G-A.C-A-C.G-C-G.G-U-C.C-G-C.U-A-U.A-A-C.G-A-G.
   Ala - Leu - Leu - Val - Arg - Asp - Gln - Cys - Ala - Ala - Trp - Ile - Arg - His - Ala - Val - Arg - Tyr - Asn - Glu -
                                                                                                                       180
2304
   U-C-A.U-A-U.G-A-A.U-U-U.A-G-G.C-U-C.G-U-U.G-U-A.G-G-G.A-A-C.G-G-A.G-U-G.U-U-U.A-C-A.G-U-U.C-C-G.A-A-G.A-A-U.A-A-U.A-A-A.
   Ser - Tyr - Glu - Phe - Arg - Leu - Val - Val - Gly - Asn - Gly - Val - Phe - Thr - Val - Pro - Lys - Asn - Asn - Lys -
                                                                                                                       200
```

2364

A-U-A.G-A-U.C-G-G.G-C-U.G-C-C.U-G-U.A-A-G.G-A-G.C-C-U.G-A-U.A-U-G.A-A-U.A-U-G.U-A-C.C-U-C.C-A-G.A-A-A.G-G-G.G-U-C.G-G-U.

Ile - Asp - Arg - Ala - Ala - Cys - Lys - Glu - Pro - Asp - Met - Asn - Met - Tyr - Leu - Gln - Lys - Gly - Val - Gly -
220

2424

G-C-U.U-U-C.A-U-C.A-G-A.C-G-C.C-G-G.C-U-C.A-A-A.U-C-C.G-U-U.G-G-U.A-U-A.G-A-C.C-U-G.A-A-U.G-A-U.C-A-A.U-C-G.A-U-C.A-A-C.

Ala - Phe - Ile - Arg - Arg - Arg - Leu - Lys - Ser - Val - Gly - Ile - Asp - Leu - Asn - Asp - Gln - Ser - Ile - Asn -
240

2484

C-A-G.C-G-U.C-U-G.G-C-U.C-A-G.C-A-G.G-G-C.A-G-C.G-U-A.G-A-U.G-G-U.U-C-G.C-U-U.G-C-G.A-C-G.A-U-A.G-A-C.U-U-A.U-C-G.U-C-U.

Gln - Arg - Leu - Ala - Gln - Gln - Gly - Ser - Val - Asp - Gly - Ser - Leu - Ala - Thr - Ile - Asp - Leu - Ser - Ser -
260

2544

G-C-A.U-C-C.G-A-U.U-C-C.A-U-C.U-C-C.G-A-U.C-G-C.C-U-G.G-U-G.U-G-G.A-G-U.U-U-U.C-U-C.C-C-A.C-C-U.G-A-G.C-U-A.U-A-U.U-C-A.

Ala - Ser - Asp - Ser - Ile - Ser - Asp - Arg - Leu - Val - Trp - Ser - Phe - Leu - Pro - Pro - Glu - Leu - Tyr - Ser -
280

2604

U-A-U.C-U-C.G-A-U.C-G-U.A-U-C.C-G-C.U-C-A.C-A-C.U-A-C.G-G-A.A-U-C.G-U-A.G-A-U.G-G-C.G-A-G.A-C-G.A-U-A.C-G-A.U-G-G.G-A-A.

Tyr - Leu - Asp - Arg - Ile - Arg - Ser - His - Tyr - Gly - Ile - Val - Asp - Gly - Glu - Thr - Ile - Arg - Trp - Glu -
300

2664

C-U-A.U-U-U.U-C-C.A-C-A.A-U-G.G-G-A.A-A-U.G-G-G.U-U-C.A-C-A.U-U-U.G-A-G.C-U-A.G-A-G.U-C-C.A-U-G.A-U-A.U-U-C.U-G-G.G-C-A.

Leu - Phe - Ser - Thr - Met - Gly - Asn - Gly - Phe - Thr - Phe - Glu - Leu - Glu - Ser - Met - Ile - Phe - Trp - Ala -
320

2724

A-U-A.G-U-C.A-A-A.G-C-G.A-C-C.C-A-A.A-U-C.C-A-U.U-U-U.G-G-U.A-A-C.G-C-C.G-G-A.A-C-C.A-U-A.G-G-C.A-U-C.U-A-C.G-G-G.G-A-C.

Ile - Val - Lys - Ala - Thr - Gln - Ile - His - Phe - Gly - Asn - Ala - Gly - Thr - Ile - Gly - Ile - Tyr - Gly - Asp -
340

2784

G-A-U.A-U-U.A-U-A.U-G-U.C-C-C.A-G-U.G-A-G.A-U-U.G-C-A.C-C-C.C-G-U.G-U-G.C-U-A.G-A-G.G-C-A.C-U-U.G-C-C.U-A-C.U-A-C.G-G-U.

Asp - Ile - Ile - Cys - Pro - Ser - Glu - Ile - Ala - Pro - Arg - Val - Leu - Glu - Ala - Leu - Ala - Tyr - Tyr - Gly -
360

2844

U-U-U.A-A-A.C-C-G.A-A-U.C-U-U.C-G-U.A-A-A.A-C-G.U-U-C.G-U-G.U-C-C.G-G-G.C-U-C.U-U-U.C-G-C.G-A-G.A-G-C.U-G-C.G-G-C.G-C-G.

Phe - Lys - Pro - Asn - Leu - Arg - Lys - Thr - Phe - Val - Ser - Gly - Leu - Phe - Arg - Glu - Ser - Cys - Gly - Ala -
380

2904

C-A-C.U-U-U.U-A-C.C-G-U.G-G-U.G-U-C.G-A-U.G-U-C.A-A-A.C-C-G.U-U-U.U-A-C.A-U-C.A-A-G.A-A-A.C-C-U.G-U-U.G-A-C.A-A-U.C-U-C.

His - Phe - Tyr - Arg - Gly - Val - Asp - Val - Lys - Pro - Phe - Tyr - Ile - Lys - Lys - Pro - Val - Asp - Asn - Leu -
400

2964

U-U-C.G-C-C.C-U-G.A-U-G.C-U-G.A-U-A.U-U-A.A-A-U.C-G-G.C-U-A.C-G-G.G-G-U.U-G-G.G-G-A.G-U-U.G-U-C.G-G-A.G-G-U.A-U-G.U-C-A.

Phe - Ala - Leu - Met - Leu - Ile - Leu - Asn - Arg - Leu - Arg - Gly - Trp - Gly - Val - Val - Gly - Gly - Met - Ser -
420

3024

G-A-U.C-C-A.C-G-C.C-U-C.U-A-U.A-A-G.G-U-G.U-G-G.G-U-A.C-G-G.C-U-C.U-C-C.U-C-C.C-A-G.G-U-G.C-C-U.U-C-G.A-U-G.U-U-C.U-U-C.

Asp - Pro - Arg - Leu - Tyr - Lys - Val - Trp - Val - Arg - Leu - Ser - Ser - Gln - Val - Pro - Ser - Met - Phe - Phe -
440

3084

G-G-U.G-G-G.A-C-G.G-A-C.C-U-C.G-C-U.G-C-C.G-A-C.U-A-C.U-A-C.G-U-A.G-U-C.A-G-C.C-C-G.C-C-U.A-C-G.G-C-A.G-U-C.U-C-G.G-U-A.

Gly - Gly - Thr - Asp - Leu - Ala - Ala - Asp - Tyr - Tyr - Val - Val - Ser - Pro - Pro - Thr - Ala - Val - Ser - Val -
460

3144

U-A-C.A-C-C.A-A-G.A-C-U.C-C-G.U-A-C.G-G-G.C-G-G.C-U-G.C-U-C.G-C-G.G-A-U.A-C-C.C-G-U.A-C-C.U-C-G.G-G-U.U-U-C.C-G-U.C-U-U.

Tyr - Thr - Lys - Thr - Pro - Tyr - Gly - Arg - Leu - Leu - Ala - Asp - Thr - Arg - Thr - Ser - Gly - Phe - Arg - Leu -
480

3204

G-C-U.C-G-U.A-U-C.G-C-U.C-G-A.G-A-A.C-G-C.A-A-G.U-U-C.U-U-C.A-G-C.G-A-A.A-A-G.C-A-C.G-A-C.A-G-U.G-G-U.C-G-C.U-A-C.A-U-A.

Ala - Arg - Ile - Ala - Arg - Glu - Arg - Lys - Phe - Phe - Ser - Glu - Lys - His - Asp - Ser - Gly - Arg - Tyr - Ile -
500

3264

G-C-G.U-G-G.U-U-C.C-A-U.A-C-U.G-G-A.G-G-U.G-A-A.A-U-C.A-C-C.G-A-C.A-G-C.A-U-G.A-A-G.U-C-C.G-C-C.G-G-C.G-U-G.C-G-C.G-U-U.

Ala - Trp - Phe - His - Thr - Gly - Gly - Glu - Ile - Thr - Asp - Ser - Met - Lys - Ser - Ala - Gly - Val - Arg - Val -
520

3324

A-U-A.C-G-C.A-C-U.U-C-G.G-A-G.U-G-G.C-U-A.A-C-G.C-C-G.G-U-U.C-C-C.A-C-A.U-U-C.C-C-U.C-A-G.G-A-G.U-G-U.G-G-G.C-C-A.G-C-G.

Ile - Arg - Thr - Ser - Glu - Trp - Leu - Thr - Pro - Val - Pro - Thr - Phe - Pro - Gln - Glu - Cys - Gly - Pro - Ala -
540

3384

A-G-C.U-C-U.C-C-U.C-G-G U-A-G

Ser - Ser - Pro - Arg
544

latter occur with a frequency not higher than expected for a random distribution. Finally the reading ends with the terminating U–A–G, as reported before[18]. This assignment is confirmed by studies of specific suppression in an *in vitro* translation system[43]. The replicase gene is followed by a 174-nucleotide untranslated, 3′-terminal segment (including U–A–G); this is shown in a more compact model in Fig. 2.

A heat stable host factor, HF (a hexamer of 72,000 daltons), is required in the Qβ-replicase reaction, but only when Qβ-RNA plus strand is the template[3]. HF binds to two regions in Qβ RNA, one of which is close to the 3′ end[44]. Although it is doubtful that HF is involved in the replication of the RNA of the MS2–R17–f2 group[45], it can nevertheless specifically bind to a single region in these viral RNAs. Senear and Steitz[44] found that HF protects a small segment in R17 RNA against nuclease, which has the following nucleotide sequence: . . . A–A–G–A–A–U–A–A–U–A–A–A–A–U–A–G . . .; undoubtedly this corresponds to the residues 2352–2367 which are located in the middle of the replicase gene. Presumably this interaction has no functional significance; HF is known to have a high affinity for poly(A)[46], and not only is the protected sequence A rich, but the T_1-oligonucleotide A–A–U–A–A–U–A–A–A–A–U–A–Gp has peculiar and very exceptional properties (strongly sticking to cellulose acetate and DEAE paper).

a

	U	C	A	G	
U	Phe 12, 16; Leu 8, (5)	Ser 7, 12, 6, 10	Tyr (5), 16; Ochre; Amber o	Cys 5, 2; Opal; Trp 9	U C A G
C	Leu 7, 15, 8, (7)	Pro 10, 4, 3, 9	His (4), (6); Gln 7, 8	Arg 11, 13, (4), (8)	U C A G
A	Ile 7, 13, (12); Met ⊙+9	Thr 4, 12, (8), 7	Asn 11, 7; Lys 9, 16	Ser (4), 9; Arg (4), (2)	U C A G
G	Val 9, 10, 6, 7	Ala 15, 7, 8, 12	Asp 19, 14; Glu 9, 13	Gly 19, 7, 8, 10	U C A G

b

	U	C	A	G	
U	Phe 19, 29; Leu 17, 11	Ser 15, 20, 16, 22	Tyr 9, 32; Ochre 1; Amber 2	Cys 6, 6; Opal; Trp 23	U C A G
C	Leu 15, 26, 15, 9	Pro 17, 10, 9, 13	His 6, 9; Gln 17, 22	Arg 21, 20, 10, 11	U C A G
A	Ile 12, 25, 19; Met 2+18	Thr 19, 21, 13, 14	Asn 17, 28; Lys 19, 26	Ser 8, 16; Arg 7, 6	U C A G
G	Val 21, 21, 16, 1+18	Ala 26, 21, 21, 23	Asp 28, 22; Glu 16, 28	Gly 37, 16, 12, 16	U C A G

Fig. 5 Code words used in the replicase gene and in the whole genome. The numbers refer to the frequency with which each code word is used. *a*, Codewords used in the replicase gene. Circled numbers indicate codons not used in the coat gene. *b*, Summational table of codewords used in the three viral genes. The initiation codewords (one G–U–G and two A–U–Gs) are counted separately.

Use of code words

The code words used in the replicase gene are summarised in Fig. 5. On what basis is the choice made between degenerate codons for a given amino acid? In some cases it may be a historical accident, like the choice between a U–C–X or an A–G–Py codon for serine. It is unlikely that there are no constraints on the third letters. Indeed, although the mutation rate is very high (as evidenced by, for example, forward and backward mutation rates), we have been able to work for at least 6 yr with an essentially unaltered sequence. An obvious constraint is that third letters may be involved in optimising the secondary and tertiary structure. Secondary structure can be maximised by a proper choice of third letters and by bringing complementary segments into proper register[47]. Although some evidence in this direction was found[13,14,48,49] in the case of the coat gene, it could not be substantiated further as more data became available[11]. We believe that relatively few, selected third letters are sufficient to bring the molecule from a random secondary structure (50–60% base pairing[49,50]; but this random structure is less stable[51]) to the level of a stable conformation as observed in the viral RNAs (73%±5 base pairing[31]). Statistical methods may not be adequate to reveal these effects convincingly. Moreover, we have no way of assessing the role of third letters in tertiary structure interactions. Neither should we forget that selection may also operate at the level of the negative strand (again no regions should be created which may function in ribosome interaction, or as nuclease targets or as binding sites for encapsidation and so on).

But it is very likely that at least for some amino acids other factors directly related to translation in *E. coli* have a predominant role in the choice of the third letter. For example, the code word G–G–U is definitely preferred for glycine and this is true for the whole genome (significant at the 0.1% level as tested by χ^2 analysis). This effect is specific and not due to a general preference for U in third positions, as found in the φX174 gene G[52].

We have suggested that isoleucine, tyrosine and/or arginine have modulating codons. Indeed, in the coat protein the codon U–A–U is not used for tyrosine, and we now conclude that for the whole genome U–A–C is definitely preferred (0.1% significance). Similarly, the code words C–G–Pu and A–G–Pu were not used for arginine in the coat gene, and it now seems that for the whole genome the preference is clearly C–G–Py > C–G–Pu > A–G–Pu (0.1% significance). These modulating triplets may be absolutely unacceptable in a gene like that for the coat, which is translated at high frequency, but can be tolerated in moderate numbers in genes whose elongation rate of translation is average. Isoleucine codons may constitute another class of modulating triplets. A–U–A is absent from the coat gene (5% significance; the coat contains only eight isoleucine residues), and is known to be recognised by a specific tRNA which is present in very small amounts in *E.coli*[53]. The A protein and the replicase genes contain several A–U–A codons, and these may very well modulate (brake) the speed of translation; in fact, in these two genes it seems that it is the A–U–U codon which is selected against (1% significance).

In summary, we have established the primary nucleotide sequence of bacteriophage MS2 RNA. The polynucleotide

chain contains 3,569 residues, 10.2% of which constitute untranslated segments. As all the genetic information is derived from this RNA molecule, the primary structure of all the virus-specified products (A protein, coat protein and replicase subunit) has also been deduced. MS2 is therefore the first living organism for which the entire primary chemical structure has been elucidated. We propose tentative models for the secondary folding of the viral RNA; parts of these are based on experimental evidence, and some aspects provide plausible bases for the explanation of biological effects. The secondary structure of the coat gene resembles a flower[14], and there are similar foldings in other parts of the molecule; the secondary structure of the whole viral RNA therefore constitutes a bouquet.

We thank R. Thijs and M. Bensch for technical assistance. This work was supported by a grant from the Belgian Fonds voor Kollektief Fundamenteel Onderzoek.

Received December 24, 1975; accepted February 9, 1976.

1 Zinder, N., (ed.) *RNA Phage Book*, 4, 353–396 (Cold Spring Harbor Laboratory, New York, 1975).
2 Argetsinger-Steitz, J., *J. molec. Biol.*, 33, 937–945 (1968).
3 Weissmann, C., Billeter, M. A., Goodman, H. M., Hindley, J., and Weber, H., *A. Rev. Biochem.*, 42, 303–328 (1973).
4 Fedoroff, N. V., and Zinder, N. D., *Proc. natn. Acad. Sci. U.S.A.*, 68, 1838–1843 (1971).
5 Haruna, I., Nozu, K., Ohtaka, Y., and Spiegelman, S., *Proc. natn. Acad. Sci. U.S.A.*, 50, 905–911 (1963).
6 Weissmann, C., Simon, L., and Ochoa, S., *Proc. natn. Acad. Sci. U.S.A.*, 49, 407–414 (1963).
7 Kamen, R., *Nature*, 228, 527–533 (1970).
8 Kondo, M., Gallerani, R., and Weissmann, C., *Nature*, 228, 525–527 (1970).
9 Wahba, A. J., *et al.*, *J. biol. Chem.*, 249, 3314–3316 (1974).
10 Kozak, M., and Nathans, D., *Bact. Rev.*, 36, 109–134 (1972).
11 Fiers, W., *et al.*, *Nature*, 256, 273–278 (1975).
12 Kolakofsky, D., and Weissmann, C., *Nature new Biol.*, 231, 42–46 (1971).
13 Fiers, W., *et al.*, *Biochimie*, 53, 495–506 (1971).
14 Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W., *Nature*, 237, 82–88 (1972).
15 De Wachter, R., Merregaert, J., Vandenberghe, A., Contreras, R., and Fiers, W., *Eur. J. Biochem.*, 22, 400–414 (1971).
16 Fiers, W., in *RNA Phage Book* (edit. by Zinder, N.), 353–396 (Cold Spring Harbor Laboratory, New York, 1975).
17 Nichols, J. L., *Nature*, 225, 147–151 (1970).
18 Vandenberghe, A., Min Jou, W., and Fiers, W., *Proc. natn. Acad. Sci. U.S.A.*, 72, 2559–2562 (1975).
19 Contreras, R., Vandenberghe, A., Volckaert, G., Min Jou, W., and Fiers, W., *FEBS Lett.*, 24, 339–342 (1972).
20 Rensing, U. F. E., Coulson, A., and Thompson, E. O. P., *Biochem. J.*, 131, 605–610 (1973).
21 De Wachter, R., and Fiers, W., *Analyt. Biochem.*, 49, 184–197 (1972).
22 Sanger, F., Brownlee, G. G., and Barrell, B. G., *J. molec. Biol.*, 13, 373–398 (1965).
23 Brownlee, G. G., *Determination of Sequences in RNA* (edit. by Work, T. S., and Work, E.) (North-Holland, Amsterdam, 1972).
24 Haegeman, G., and Fiers, W., *Eur. J. Biochem.*, 36, 135–143 (1973).
25 Volckaert, G., and Fiers, W., *Analyt. Biochem.*, 62, 573–583 (1974).
26 Contreras, R., and Fiers, W., *Analyt. Biochem.*, 67, 319–326 (1975).
27 Contreras, R., and Fiers, W., *FEBS Lett.*, 16, 281–283 (1971).
28 Volckaert, G., Min Jou, W., and Fiers, W., *Analyt. Biochem.* (in the press).
29 Tinoco, I., Jr, *et al.*, *Nature new biol.*, 246, 40–41 (1973).
30 Borer, P. N., Dengler, B., and Tinoco, I., Jr, *J. molec. Biol.*, 86, 843–853 (1974).
31 Boedtker, H., *Biochemistry*, 6, 2718–2727 (1967).
32 Slegers, H., Clauwaert, J., and Fiers, W., *Biopolymers*, 12, 2033–2044 (1973).
33 Gralla, J., Steitz, J. A., and Crothers, D. M., *Nature*, 248, 204–208 (1974).
34 Klug, A., Ladner, J., and Robertus, J. D., *J. molec. Biol.*, 89, 511–516 (1974).
35 Lodish, H. F., *Nature*, 220, 345–349 (1968).
36 Osborn, M., Weber, K., and Lodish, H. F., *Biochem. biophys. Res. Commun.*, 41, 748–756 (1970).
37 Argetsinger-Steitz, J., *Nature*, 224, 957–964 (1969).
38 Shine, J., and Dalgarno, L., *Nature*, 254, 34–38 (1975).
39 Argetsinger-Steitz, J., and Jakes, K., *Proc. natn. Acad. Sci. U.S.A.*, 72, 4734–4738 (1975).
40 Lodish, H. F., *J. molec. Biol.*, 50, 689–702 (1970).
41 Argetsinger-Steitz, J., *J. molec. Biol.*, 73, 1–16 (1973).
42 Bernardi, A., and Spahr, P. F., *Proc. natn. Acad. Sci. U.S.A.*, 69, 3033–3037 (1972).
43 Atkins, J. F., and Gesteland, R. F., *Molec. gen. Genet.*, 139, 19–31 (1975).
44 Senear, A. W., and Argetsinger-Steitz, J., *J. biol. Chem.* (in the press).
45 Fedoroff, N. V., and Zinder, N. D., *Nature new Biol.*, 241, 105–108 (1973).
46 Carmichael, G. G., Weber, K., Niveleau, A., and Wahba, A. J., *J. biol. Chem.*, 250, 3607–3612 (1975).
47 Min Jou, W., Haegeman, G., and Fiers, W., *FEBS Lett.*, 13, 105–109 (1971).
48 Adams, J. M., Jeppesen, P. G. N., Sanger, F., and Barrell, B. G., *Nature*, 223, 1009–1014 (1969).
49 Fitch, W. M., *J. molec. Biol.*, 3, 279–291 (1974).
50 Gralla, J., and Delisi, C., *Nature*, 248, 330–332 (1974).
51 Ricard, B., and Salser, W., *Biochem. biophys. Res. Commun.*, 63, 548–554 (1975).
52 Air, G. M., Blackburn, E. H., Sanger, F., and Coulson, A. R., *J. molec. Biol.*, 96, 703–719 (1975).
53 Harada, F., and Nishimura, S., *Biochemistry*, 13, 300–307 (1974).
54 Fiers, W., *et al.*, *Abstr. 3rd int. Cong. Virol., Madrid*, 17 (1975).
55 Weiner, A. M., and Weber, K., *Nature new Biol.*, 234, 206–209 (1971).

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset