(Reprinted from Nature, Vol. 237, No. 5350, pp. 82–88, May 12, 1972)

Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein

W. MIN JOU, G. HAEGEMAN, M. YSEBAERT & W. FIERS

Laboratory of Molecular Biology and Laboratory of Physiological Chemistry, State University of Ghent, Belgium

By characterization of fragments, isolated from a nuclease digest of MS2 RNA, the entire nucleotide sequence of the coat gene was established. A "flower"-like model is proposed for the secondary structure. The genetic code makes use of 49 different codons to specify the sequence of the 129 amino-acids long coat polypeptide.

DETAILED knowledge of the structure of a viral gene should provide a firmer basis for explaining the mechanism, specificity and regulation of translation. Because the genes of RNA bacteriophages are part of the viral genome, their structures may also have evolved for optimal functioning in replication and encapsulation.

The RNA of the closely related phages f2, MS2 and R17, and especially the cistron coding for the coat protein, has been used extensively for *in vitro* study of protein synthesis[1]. ^{32}P-labelled bacteriophage MS2 RNA was promising as a material for investigating the primary structure of this viral messenger[2]. By using limited enzymatic digestion of the RNA at low temperatures and separating the products on polyacrylamide gels, we isolated RNA-fragments of limited length, suitable for direct sequence analysis[3]. Sanger *et al.*[4] were the first to link nucleotide sequences with specific functions in translation, isolating a "hairpin", 57 nucleotides long, from the R17 coat protein cistron, from which they could deduce a set of codons used in the genetic message. Argetsinger-Steitz[5] identified a hairpin containing the initiating AUG-codon, while Nichols[6] determined the nucleotide sequence of another hairpin, which contained the coat termination signal; information on two more hairpin-regions was also reported[7]. Working with the phage MS2, we identified and characterized similar coat cistron hairpins[8]. These sequences were extended[9], and we can now present the entire sequence of the coat protein gene.

Pure Fragments from the Coat Gene

All RNA fragments used for sequencing the coat protein cistron originated from ribonuclease T_1 hydrolysis of uniformly labelled ^{32}P-MS2 RNA, and subsequent separation on neutral polyacrylamide gels. Two sets of conditions were used: 1° hydrolysis with 1 U of enzyme per 20 μg of RNA for 30 min at 0° C and separation on a 12% gel; and 2° with 1 U per 100 to 200 μg of RNA for 30 min at 0° C, followed by separation on a 6% gel. Further details and typical patterns of both separations have been published before[10,11].

As the amino-acid sequence of the coat polypeptide is known[12–15] (albeit incorrectly as it turned out), a nucleotide sequence can be deduced for this gene, leaving open the many

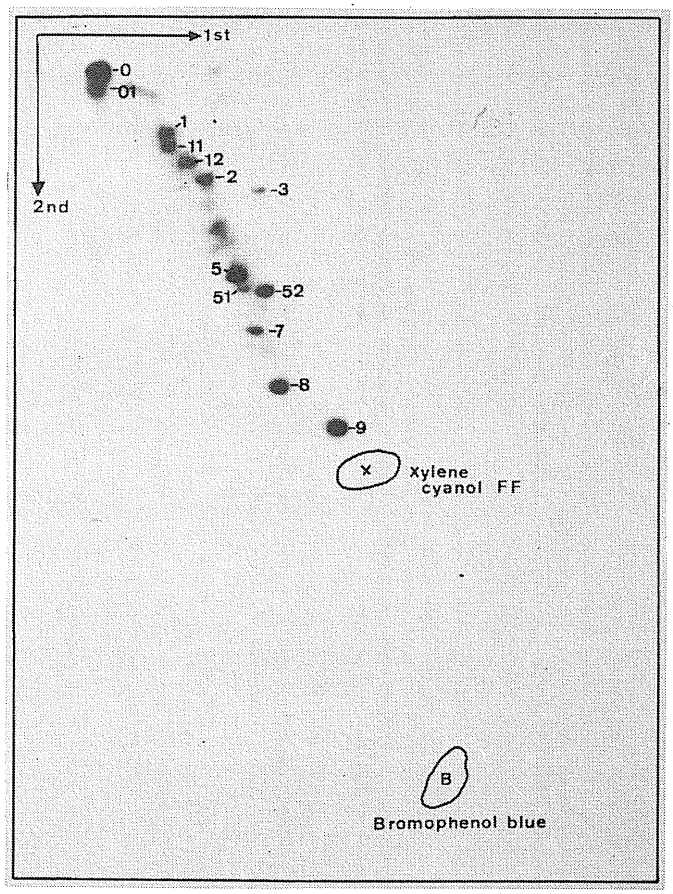

Fig. 1 Two-dimensional separation of band β4. The RNA of band β4 was extracted and subjected to two-dimensional gel electrophoresis. The arrows indicate the directions of the run: the first dimension was on an 8% gel in 0.025 M citric acid+6 M urea; subsequent separation was on a 16% gel in 0.04 M Tris-acetate, *p*H 8. The positions of the dye markers xylene cyanol FF (X) and bromophenol blue (B) are marked on the figure. Spots *b*1, *b*2 and *b*5 correspond to different regions of the coat protein cistron (Table 2). Spot *b*12 of this particular separation is a rarely observed fragment starting around the GGCGGA sequence (coding for amino-acids 13–14) and going till arrow 14 in Fig. 2.

positions which are not or only partly fixed due to the degeneracy of the genetic code. The information was sufficient to screen the relatively long oligonucleotides, obtained by complete ribonuclease T_1-hydrolysis of the gel bands and fingerprinting, and to see whether they could be derived from the coat gene. This resulted in isolation and sequencing of five fragments spread over the cistron[8]. A similar approach was used to detect regions from the coat protein cistron of R17[4,6,7] and f2 RNA[16].

Seeking longer fragments we examined slower moving bands in the 12% gel. But more important was a thorough analysis of the bands derived from the 6% gel (under the milder digestion conditions, known sequences were present in fragments of greater chain length). Each band was subjected to two-dimensional polyacrylamide gel electrophoresis[17], involving an acidic gel system in the first dimension[10] (10% or 8% gel), followed by a separation at neutral *p*H (20% or 16% gel respectively). Such a separation is shown in Fig. 1. After systematic T_1 mapping of all spots present in reasonable amounts, many fragments were recognized as being derived from the coat protein gene (Table 1).

Primary Structure Determination

We mainly used the sequencing methods developed by Sanger *et al.*[4,19,20] for the analysis of purified fragments; details of these sequence determination studies will be published elsewhere. We could not, however, build up the entire nucleotide sequence of the gene in this way. We ended up with four non-overlapping regions between the arrows 0–4, 4–14, 14–22 and 22–28 respectively (Table 1, Figs. 2 and 3). Although there is speculative evidence for their relative order we cannot exclude, purely on the basis of nucleotide sequence data, the possibility of extra nucleotides at three points (arrows 4, 14 and 22). The direct linkage is mainly based on the amino-acid sequence of the MS2 phage[14,15], which has recently been completed (Vandekerckhove and Van Montagu, personal communication). The nucleotide sequence obtained is shown in Fig. 4, together with the amino-acid sequence for which it codes. It comprises the ribosomal binding site for the coat protein cistron, the whole translated region, an intercistronic stretch of 36 nucleotides, and the ribosome attachment site of the RNA-polymerase gene[5].

Before discussing other aspects, a few additional comments on the primary structure determination are appropriate. So far, the exact order of the T_1 oligonucleotides in the region between arrows 3 and 4 (coding for the amino-acids 7–18) has not yet been determined, since fragments containing this sequence were difficult to obtain in sufficient quantity. Nevertheless, the results are not compatible with the amino-acid sequence reported for the corresponding region in the f2 and R17 proteins[12,13]. By introducing three amino-acid changes, the nucleotide sequence data can be explained. These are situ-

Table 1 RNA Fragments from MS2 Coat Protein Gene isolated from Limited Ribonuclease T_1 Hydrolysates

Fragment	Located between arrows No.		Chain length	Estimation of chain length		Yield
				12% gel	6% gel	
D6*b*11	22	24	30	26–31		(Irregularly found)
D5*b*5	2	3	31–32	29–35		Good
D5*b*3	5	7	30	29–35		Variable
D1*z*2; δ5*b*0	24	25	43	40–50	40–54	Good (irregularly found)
δ4*b*2	0	3	±52		46–52	Variable
C7*z*1	18	19	57	46–57		Moderate
C6*t*1; δ4*b*2	13	14	32	49–61	46–62	Good; variable
C6*t*2; δ4*b*7	14	15	27	49–61	46–62	Good; variable
C5*z*1; δ3*b*2	8	10	61	52–64	51–66	Low (irregularly found)
C5*z*3	8	9	34	52–64		Good
C5*z*5	9	10	27	52–64		Good
δ2*b*7	1	4	78		63–82	Variable
δ2*b*4	18	20	72		63–82	(Irregularly found)
δ2*b*1	24	27	71		63–82	(Irregularly found)
C2*b*3	16 (15)	20	85 (87)	68–86		Variable
δ1*b*106	14	17	36		72–94	(Irregularly found)
δ1*b*2	22 (23)	25	73 (69–70)		72–94	Moderate
γ8*b*4	0	4	±88		86–112	Variable
B13*b*1; γ8*b*6	15	21	98	85–110	86–112	Moderate; good
γ8*b*3	22	26	89		86–112	Variable
γ6*b*1	22	27	101		100–132	(Irregularly found)
B4*b*5	14	20	114	128–172		(Irregularly found)
B2*b*23; γ4*b*168	12	14	37	150–200	150–200	Moderate
B2*b*22; γ4*b*16	11	14	40	150–200	150–200	Low
B2*b*11; γ4*b*4	14	21	125	150–200	150–200	Moderate
γ2*b*13	12	14	37		190–260	(Irregularly found)
γ2*b*4	14	22	141		190–260	(Irregularly found)
γ1*b*11	9	14	69		212–280	Good
γ1*b*7	14	21	125		212–280	Good
β7*b*14	9	14	69		240–330	Good
β7*b*7	14	21	125		240–330	Good
β7*b*6	14	22	141		240–330	Moderate
A3*b*7	14	21	125	245–335		(Irregularly found)
β5*b*91	8	14	103		295–400	(Irregularly found)
β5*b*52	14	21	125		295–400	(Irregularly found)
β4*b*4	6	14	138		340–450	Variable
β4*b*2	4	14	145		340–450	Good
β4*b*5	14	21	125		340–450	Good
β4*b*6	14	22	141		340–450	Variable
β4*b*1	22	28	±180		340–450	Good

The nomenclature of the gel bands is described in reference 11. Bands from the 12% gel are referred to by Roman capitals, while Greek letters are used for bands of the 6% gel. Fragments already described earlier[8] have been included for completeness; they were purified by homochromatography on thin layers[18] (indicated as *t*), or by acidic gel electrophoresis[10] (indicated as *z*). All other components were obtained after two-dimensional gel electrophoresis (indicated as *b*). The numbers of the arrows correspond to Fig. 2. Fragments taking part in a complex have been connected with hyphens, but not all components of such complexes are necessarily listed. The same components as found in β4 were also present in several slower moving bands (for example, β3, α3, α2 and α1).

ated at positions 11 (Asp) and 12 (Asn) and 17 (Asp). Each difference can be accounted for by a single A↔G transition in the codons AAPyp (Asn)←→GAPyp (Asp). Recent amino-acid sequence studies of the MS2 coat protein (Vandekerckhove and Van Montagu, personal communication) have directly confirmed the changes which we propose. The coat protein of MS2 thus differs in three positions from R17 and in a fourth place from f2 (this being the only difference between R17 and f2). This agrees well with the serological data of Scott[21], who showed that although MS2–R17–f2 were closely related, the former two were nevertheless not identical. We found this difficult to explain as their coat proteins, the most obvious candidate as antigenic determinant, were thought to be identical.

Repeated analysis of the fragment 14–22 revealed the presence of a heterogeneity in the region coding for amino-acid 109 (Gln). We found that in our RNA population both CAA and CAG codons were used (in a ratio of 2 to 1). Fragments containing the different sequences were sometimes separated (albeit incompletely) on the two-dimensional gel. Two T_1-maps of this sequence (where the alternatives appear as CAAG and CAG) are shown in Fig. 4. After returning to our original MS2 stock for phage growth[2], however, only CAA was found in that position.

Secondary and Tertiary Structure

Physico-chemical studies had revealed that phage RNAs contain a high degree of secondary structure[22,23]. The helicity was estimated as 63% to 82%[24–26]. It was first shown for ribosomal RNA[27,28] and subsequently for phage RNAs[3,4,11] that limited treatment with nucleases at low temperature results in a characteristic non-random pattern of breakdown. This is generally believed to be indicative of a specific secondary and tertiary structure, although the presence of a limited number of different conformational states cannot be excluded. Alternative conformations have been described for tRNA[29] and there is some experimental evidence for phage RNA as well[30,31]. We believe, however, that our entire approach of isolating specific fragments from partial digests was successful only because the same secondary structural features were present in most if not all molecules of the population. These partial digestion products are convenient starting material for the construction of a secondary structure model.

Extensive characterization of material present in the gel bands of both primary systems provided numerous examples for the importance of specific secondary binding forces between two (or more) RNA segments. In agreement with others[32,33], our results indicate that the mobility on a neutral gel is mainly a logarithmic function of the molecular weight. For a number of fragments, there is reasonable agreement between the actual chain length, known from the primary structure, and the estimate, based on the mobility in the gel (Table 1). In other cases there is clearly a large discrepancy, which can most easily be explained by assuming that they behaved in the neutral gel system as part of a tightly-bound complex, for example, band C5 contains the fragments *z*3 ($n=34$) and *z*5 ($n=27$) but also *z*1 ($n=61$). It should be noted here that the primary gel (either 12% or 6%) is run under neutral conditions, whereas the different systems used for the subsequent fractionation make use of denaturing conditions (low *p*H and urea). Slower moving bands from the 6% gel contain even more intricate complexes. For example, β4 contains not only the fragments *b*2 and *b*5 linked by the interactions in region *f*, but also the fragment *b*1 occurs consistently in the same band. Its presence can most easily be explained by assuming that it is linked to the former two (and most likely to a segment of the sequence appearing simultaneously in the complex; Fig. 3). Together the three fragments comprise approximately 460 nucleotides, in agreement with the estimated chain length of β4 (340–450).

The "Flower" Model

It seemed reasonable to start from the partial digestion products (and complexes) for the construction of a secondary

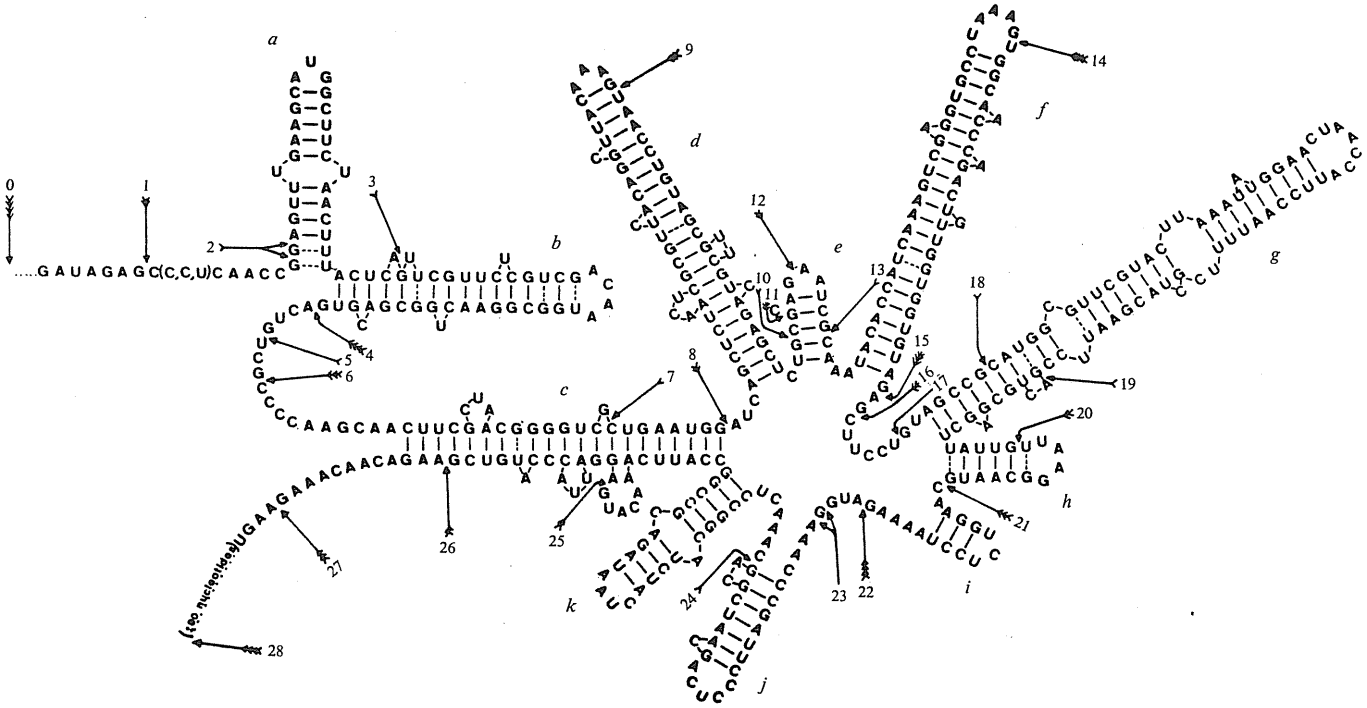

Fig. 2 A model for the secondary structure of the coat protein gene (the "Flower" model). Splitting points for T_1-ribonuclease observed in the partial digests are indicated by arrows. The number of feathers of the arrows give a semi-quantitative measure of the susceptibility of the Gp–N bond (arrows with no feathers point to occasionally found splitting points). Since a part of the sequence may exist in an alternative configuration (Fig. 6) some of the splitting points marked on this figure may in fact reflect the other structure (arrow 26). The different fragments isolated can be read from this figure in combination with the data summarized in Table 1. Base-paired regions are termed *a* to *k*. Table 2 lists the stability numbers of these regions, calculated according to Tinoco *et al.*[34].

structure model of the coat protein cistron. Beginning with the shortest fragments (Fig. 3) and optimalizing secondary structure by also allowing G : U base pairs and slight deformations such as looped-out bases (bulges) and interior loops (bubbles), one can draw a series of models, which are similar in general outline but which vary in detail. Tinoco, Uhlenbeck and Levine[34] presented a simple method which permitted the estimation of the thermodynamic stability of various pairing schemes, and thus the choice of the most stable configuration. In this method, stabilizing and destabilizing features are expressed in quantitative, algebraic terms. On this basis, the model of the coat gene presented in Fig. 2 was derived. We pro-

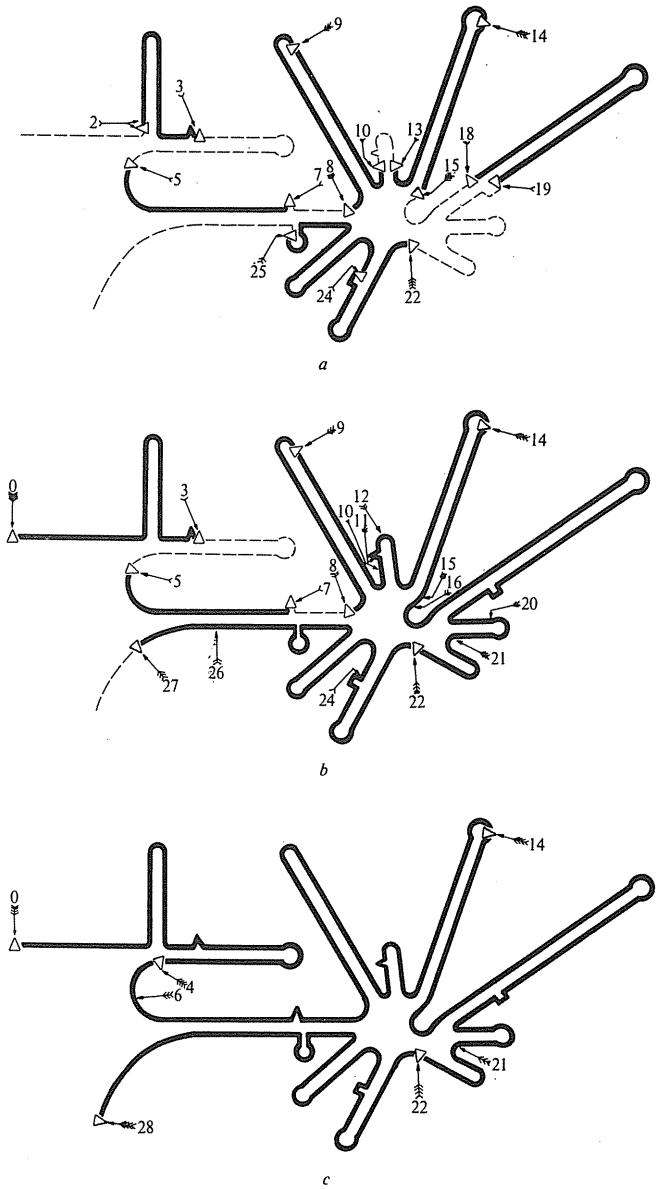

Fig. 3 Outline, illustrating different stages of hydrolysis. Fragments typical for a certain stage of breakdown are represented as a solid line. The model, and the arrows, correspond to Fig. 2. Both ends of a fragment are indicated by open triangles. Arrows not pointing to a triangle indicate positions of splitting in intermediate stages of hydrolysis. *a*, Shortest fragments isolated from different parts of the molecule (the region between arrows 5 and 7 forms a short hairpin in the alternative model, Fig. 6). *b*, Different intermediary stages of hydrolysis. The three hairpin-loops in the middle of the cistron appear as one complex. The beginning and the end of the cistron are also elongated. *c*, Finally, the whole region studied appears as four fragments. The beginning of the cistron (between arrows 0 and 4) is found independent from the others. The latter three fragments are present together, presumably as a complex, in bands β4, β3, α3, α2 and α1 (two of these fragments can occur in slightly varying length).

```
... (G)· AUA· GAG· CCC· UCA· ACC· GGA· GUU· UGA· AGC· AUG·
GCU· UCU· AAC· UUU· ACU· CAG· UUC· GUC· GAC· AAU· GGC· GGA· ACU· GGC· GAC· GUG· ACU· GUC· GCC· CCA· AGC· AAC· UUC·
Ala  Ser  Asn  Phe  Thr  Gln  Phe  Val  Asp  Asn  Gly  Gly  Thr  Gly  Asp  Val  Thr  Val  Ala  Pro  Ser  Asn  Phe
  1                   5                   10                  15                  20                  25

GCU· AAC· GGG· GUC· GCU· GAA· UGG· AUC· AGC· UCU· AAC· UCG· CGU· UCA· CAG· GCU· UAC· AAA· GUA· ACC· UGU· AGC· GUU· CGU· CAG·
Ala  Asn  Gly  Val  Ala  Glu  Trp  Ile  Ser  Ser  Asn  Ser  Arg  Ser  Gln  Ala  Tyr  Lys  Val  Thr  Cys  Ser  Val  Arg  Gln
                30                  35                  40                  45                  50

AGC· UCU· GCG· CAG· AAU· CGC· AAA· UAC· ACC· AUC· AAA· GUC· GAG· GUG· GUG· CCU· AAA· GUG· GCA· ACC· CAG· ACU· GUU· GGU· GGU· GUA·
Ser  Ser  Ala  Gln  Asn  Arg  Lys  Tyr  Thr  Ile  Lys  Val  Glu  Val  Pro  Lys  Val  Ala  Thr  Gln  Thr  Val  Gly  Gly  Val
                55                  60                  65                  70                  75

GAG· CUU· CCU· GUA· GCC· GCA· UGG· CGU· UCG· UAC· UUA· AAU· AUG· GAA· CUA· ACC· AUU· CCA· AUU· UUC· GCU· ACG· AAU· UCC· GAC·
Glu  Leu  Pro  Val  Ala  Ala  Trp  Arg  Ser  Tyr  Leu  Asn  Met  Glu  Leu  Thr  Ile  Pro  Ile  Phe  Ala  Thr  Asn  Ser  Asp
                80                  85                  90                  95                  100

UGC· GAG· CUU· AUU· GUU· AAG· GCA· AUG· CAA· GGU· CUC· CUA· AAA· GAU· GGA· AAC· CCG· AUU· CCC· UCA· GCA· AUC· GCA· GCA· AAC·
Cys  Glu  Leu  Ile  Val  Lys  Ala  Met  Gln  Gly  Leu  Leu  Lys  Asp  Gly  Asn  Pro  Ile  Pro  Ser  Ala  Ile  Ala  Ala  Asn
                105                 110                 115                 120                 125

UCC· GGC· AUC· UAC· UAA· UAG· ACG· CCG· GCC· AUU· CAA· ACA· UGA· GGA· UUA· CCC· AUG· UCG· AAG· ACA· ACA· AAG· AAG· (U)
Ser  Gly  Ile  Tyr                                                                Ser  Lys  Thr  Thr  Lys  Lys
                129                                                                 1                   5
```

Fig. 4 Nucleotide sequence of the coat protein gene, together with the amino-acid sequence it specifies. The gene is preceded and followed by untranslated intercistronic regions. The numbers refer to the position of the amino-acid residues in the coat protein (1–129) and in the polymerase molecule (1–6).

pose to call it the "flower" model, as one can discern a stalk with petals. 66.4% of the nucleotides are involved in helical regions.

The stability constants of the various base-paired segments are summarized in Table 2. No alternative structures with comparable stability could be obtained in most parts of the sequence, except for the regions between arrows 15 and 18 and between 19 and 22 (or even 24); these parts of the structure should therefore be considered as highly tentative. Another part of the sequence can clearly be represented in two different ways, either as written as in Fig. 2 (stem *c* of the figure), or alternatively as a structure with two hairpin-loops and a short stem as shown in Fig. 6. This model has the same overall stability (Table 2).

In general, the position of the splitting points (arrows in Fig. 2) confirms the proposed secondary structure. Four-feather arrows point to a G-residue in a loop at the top of a hairpin (arrow 14) or in single-stranded regions (arrows 4 and 22). Many other arrows point to G-residues which either lie in single-stranded regions or adjacent to the end of a hairpin or to looped-out bases, whereas almost every G-residue in a double-stranded region is not split by the enzyme.

In constructing the model, many choices were based on the stability constants as defined by Tinoco, Uhlenbeck and Levine[34]. As these authors point out, the theoretical evaluation of the stability of polynucleotide conformations will certainly be refined in the future. We believe, however, that these improvements will mainly involve details rather than the general outline of the proposed model, as at least a large part is experimentally supported by several lines of evidence. Another theoretical improvement will be the construction of one or more secondary structures by computer evaluation of the primary structure. This, however, should preferably be done after the entire viral nucleotide sequence is known and perhaps also after the program has been refined. In addition, experimental approaches will undoubtedly contribute to a better understanding of the polynucleotide conformation.

Finally, it should be stressed that the "flower" model, as presented in Fig. 2, is presumably further folded in an intricate, three-dimensional superstructure. This may explain why some G-residues, apparently situated in single-stranded regions, are not, or only rarely, split.

Biological Implications

J. Argetsinger-Steitz[5] determined the nucleotide sequence of the region, surrounding the initiating AUG for each of the three R17 cistrons. Part of the MS2 A-protein ribosome binding site has previously been sequenced[11]. We now find that the ribosome binding regions for the MS2 coat cistron and the MS2 RNA-polymerase cistron are virtually identical to their R17 counterparts (only a G to A change is noted in the former region; a similar base change, however, was found by Cory, Spahr and Adams[35] for R17 RNA and by Gupta *et al.*[36] for f2 RNA). So the untranslated regions, preceding the initiating AUG-codons, seem to be genetically conserved.

The secondary-structure model, however, does not provide a clear clue as to what constitutes a ribosomal binding site. The first AUG of the coat cistron is conveniently located on top of a hairpin[5]. But no such structure can be proposed for the first AUG of the RNA-polymerase cistron; even in the alternative model (Fig. 6), this AUG is buried by base-pairing. On the other hand, several AUGs seem to be located in single-stranded regions (for example, one between the hairpins *i* and *j*, and another in the untranslated region preceding the RNA-polymerase cistron). Their inability to bind ribosomes may perhaps be explained by their being buried in a three-dimensional folding.

The nucleotide triplets, which code for the 129 amino-acids long coat polypeptide, are listed in Table 3. Not unexpectedly,

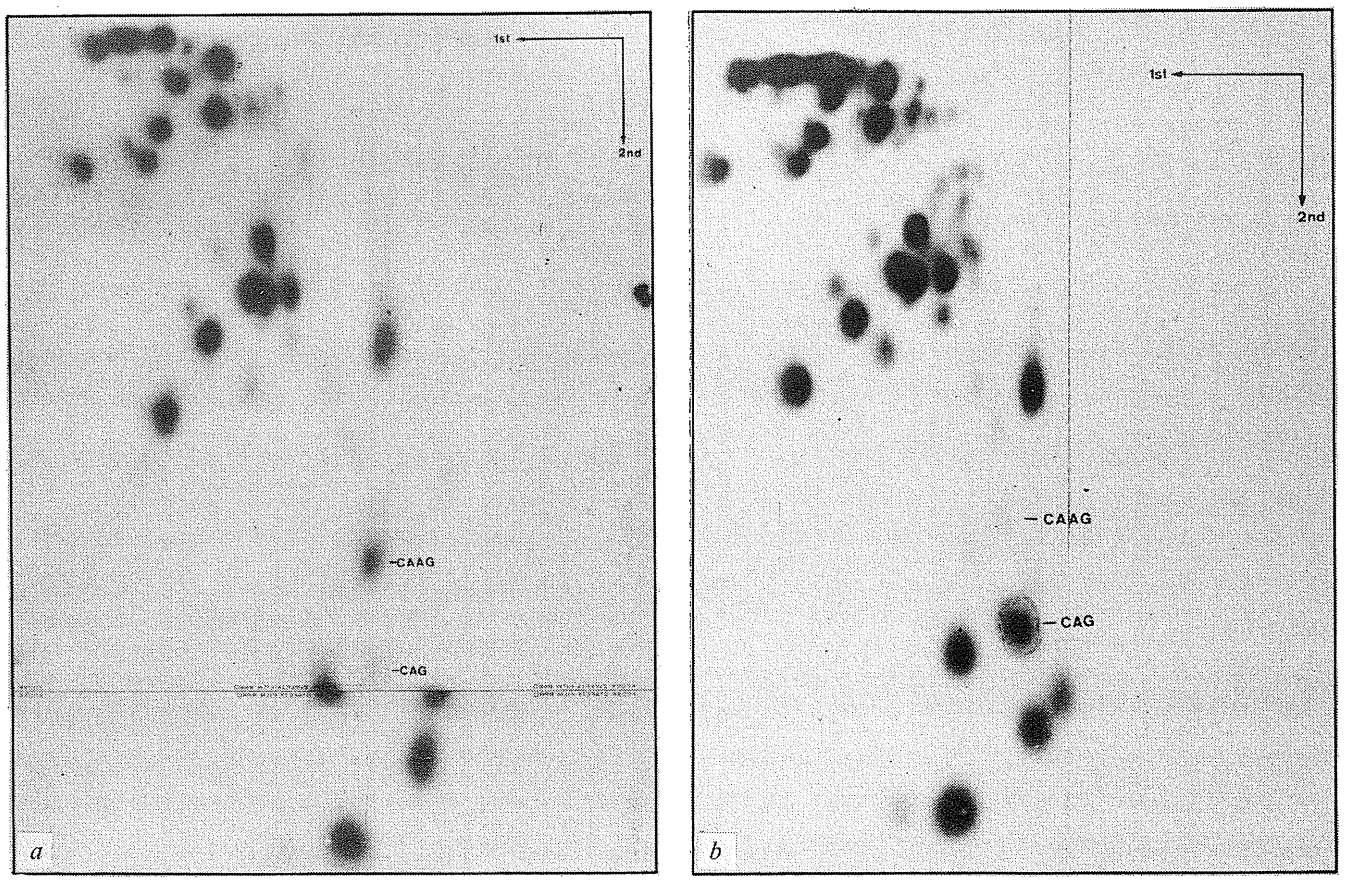

Fig. 5 Ribonuclease T_1 fingerprints of the fragment between arrows 14 and 22 (Fig. 2), illustrating the observed sequence heterogeneity in this region. *a*, Fragment in which the sequence CAA·G (coding for amino-acid 109; Gln) is predominant; *b*, sequence "variant" separated from the former by two-dimensional gel electrophoresis. Spot CAAG is replaced here almost completely by CAG (and G) (such fragments were no longer observed after returning to our original phage stock).

Table 2 Stability Numbers of Base-paired Regions

Region	Stability number
a	4
b	6
c	10
d	16
e	1
f	19
g	13
h	−1
i	0
j	1
k	5
l	4
m	3
n	3

The stability numbers were calculated by the method of Tinoco *et al.*[34]. The base-paired regions are indicated in Fig. 2 (*a–k*) and in Fig. 6 (*l–n*). The latter figure represents an alternative secondary structure model for roughly that part of the molecule involved in the base-paired region *c* (Fig. 2). Stable secondary structures have positive values, negative values characterize structures unstable with respect to the single strand (region *h* is nevertheless represented as a double-stranded structure in Fig. 2, because of our data on the nature of the fragments obtained).

our results fully confirm, in a most straightforward way, the genetic code word dictionary compiled by Ochoa–Nirenberg–Khorana. In all, 49 different code words are used. For a few amino-acids the choice between the degenerate codons seems to be non-random, although this could still be a coincidence. Several places in the code table remain unoccupied, and it seems unlikely that chance alone can explain this. Especially noteworthy is the absence of AUA as a codon for Ile and UAU as a codon for Tyr; these results are corroborated by the preliminary information on codons used in the polymerase cistron[9].

Although in some cases the choice between degenerate code words may be dictated by the properties of the *Escherichia coli* translation machinery, in other instances it may be based on secondary structure requirements for the viral RNA. In fact, there is (statistically weak) evidence supporting this assumption.

Table 3 Codons found in the MS_2 Coat Protein Gene

	U	C	A	G	
U	Phe °, °°°	Ser °°°, °°, °°, °°	Tyr °°°°	Cys °, °	U
	Leu °		Ochre ○	Opal	C
			Amber ○	Trp °°	A
					G
C	Leu °°, °°, °°	Pro °°, °, °°, °	His	Arg °°°, °	U
			Gln °, °°°°°		C
					A
					G
A	Ile °°°°, °°°°	Thr °°°°, °°°°, °	Asn °°°°, °°°°°°	Ser °°°°, °°°°	U
	Met ○°°		Lys °°°°°, °	Arg	C
					A
					G
G	Val °°°°, °°°°, °°°, °°°	Ala °°°°°, °°, °°°°°°, °	Asp °, °°°	Gly °°°, °°°, °°, °	U
			Glu °°, °°°		C
					A
					G

The Polarity Effect

It is known, both from *in vivo* and *in vitro* results, that at least part of the coat gene has to be translated in order to allow translation of the following RNA polymerase cistron[37–40]. Coat amber mutants at position 6 are strongly polar under non-permissive conditions; similar mutants at positions 50, 54 or 70, however, are not polar. A logical explanation is that polarity is relieved when ribosomes travel over the region somewhere between positions 6 and 50 of the coat cistron.

This hypothesis can now be explained in molecular detail by the "flower" model. Indeed, when ribosomes translate the region of amino-acids 24 to 32 of the coat cistron, they have to open up the hairpin region *c*. In this way they release the opposing strand, and, conceivably, this is sufficient for other ribosomes to bind at the RNA-polymerase-initiating AUG.

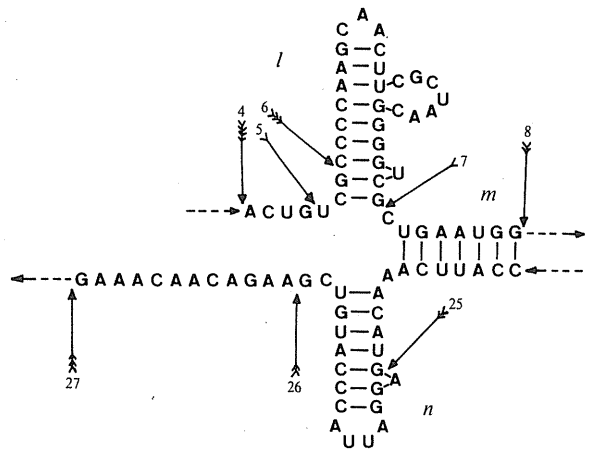

Fig. 6 Alternative model for a part of the coat protein gene and for the initiation site of the polymerase gene. Base-paired regions are termed *l* to *n*. The initiating AUG-triplet of the polymerase cistron is present in the CCC.AUG.U sequence of region *n*. For further explanation see legend to Fig. 2.

It is known that the viral RNA can adopt a conformation which does not show the polarity effect[31]. This can perhaps be explained by the alternative model, shown in Fig. 6.

Specificity of Mutagenesis

If transitions alone are legitimate, then amber codons (UAG) could arise from Gln codons (CAG) or from Trp codons (UGG). Zinder and Cooper[41] isolated a series of amber suppressor-sensitive mutants of f2 by treatment with nitrous acid. Some of these mutations were in the coat cistron, and their location was shown by amino-acid sequence analysis[42] to be either at position 6 (Gln) or position 70 (Gln). A collection of similar amber mutants from the phage R17 was isolated by using nitrous acid and fluorouracil as mutagens[38,43]. These coat mutants corresponded to positions 6, 50 or 54 (all Gln). Using nitrous acid as well as hydroxylamine, Van Montagu (refs. 44, 45 and personal communication) isolated and characterized a series of MS2 amber mutants. The nitrous acid induced coat mutants were at position 6 (the major fraction) and at position 50, while the hydroxylamine induced coat mutations were localized at positions 50, 70 and 82 (the latter is the only known Trp-mutation). It is of interest to examine these mutation data in the light of the structural model. The CAG (Gln) codons at positions 6, 50, 54 and 70, where mutations can occur, are all involved in a discontinuity of the double-stranded hairpin stem (an A or C residue is looped out). The CAG (Gln) codon at position 40, where no mutation has ever been found, forms part of an uninterrupted bihelical segment (either the mutagens do not attack

this C, or else base-pairing in this region is critical and an induced mutation in the opposite strand would give rise to an unacceptable missense, for example, a UGU Cys to UAU Tyr). The only other Gln-residue in the sequence, at position 109, is coded by a CAA-codon, which would result in a non-viable ochre mutation (although in some of our stocks a CAG-variety at this position was present). Finally, the UGG-codon which was mutated by hydroxylamine treatment to UAG is also found in a double-stranded region.

This work was supported by a grant from the Fonds voor Kollektief Fundamenteel Onderzoek. W. M. and G. H. thank the Nationaal Fonds voor Wetenschappelijk Onderzoek for fellowships. A sample of U_2-ribonuclease was supplied by Sankyo Co., Tokyo. We thank Mrs M. Borremans, F. Duerinck, A. Raeymaekers and J. Wevers for their cooperation, and Dr M. Van Montagu for discussions.

Received January 20; revised April 14, 1972.

[1] *Cold Spring Harbor Symp. Quant. Biol.*, **34**, The Mechanism of Protein Synthesis (1969).
[2] Fiers, W., Lepoutre, L., and Vandendriessche, L., *J. Mol. Biol.*, **13**, 432 (1965).
[3] Min Jou, W., Hindley, J., and Fiers, W., *Arch. Intern. Physiol. Biochem.*, **76**, 194 (1968).
[4] Adams, J. M., Jeppesen, P. G. N., Sanger, F., and Barrell, B. G., *Nature*, **223**, 1009 (1969).
[5] Steitz, J. A., *Nature*, **224**, 957 (1969).
[6] Nichols, J. L., *Nature*, **225**, 147 (1970).
[7] Jeppesen, P. G. N., Nichols, J. L., Sanger, F., and Barrell, B. G., *Cold Spring Harbor Symp. Quant. Biol.*, **35**, 13 (1970).
[8] Min Jou, W., Haegeman, G., and Fiers, W., *FEBS Lett.*, **13**, 105 (1971).
[9] Fiers, W., Contreras, R., De Wachter, R., Haegeman, G., Merregaert, J., Min Jou, W., and Vandenberghe, A., *Biochimie*, **53**, 495 (1971).
[10] De Wachter, R., and Fiers, W., *Methods in Enzymology* (edit. by Colowick, S. P., and Kaplan, N. O.), **21**, 167 (Academic Press, New York and London, 1971).
[11] De Wachter, R., Merregaert, J., Vandenberghe, A., Contreras, R., and Fiers, W., *Europ. J. Biochem.*, **22**, 400 (1971).
[12] Weber, K., and Konigsberg, W., *J. Biol. Chem.*, **242**, 3563 (1967).
[13] Weber, K., *Biochemistry*, **6**, 3144 (1967).
[14] Lin, J. Y., Tsung, C. M., and Fraenkel-Conrat, J., *J. Mol. Biol.*, **24**, 1 (1967).
[15] Vandekerckhove, J., Francq, H., and Van Montagu, M., *Arch. Intern. Physiol. Biochim.*, **77**, 175 (1969).
[16] Nichols, J. L., and Robertson, H. D., *Biochim. Biophys. Acta*, **228**, 676 (1971).
[17] De Wachter, R., and Fiers, W., *Analyt. Biochem.* (in the press).
[18] Brownlee, G. G., and Sanger, F., *Europ. J. Biochem.*, **11**, 395 (1969).
[19] Sanger, F., Brownlee, G. G., and Barrell, B. G., *J. Mol. Biol.*, **13**, 373 (1965).
[20] Brownlee, G. G., and Sanger, F., *J. Mol. Biol.*, **23**, 337 (1967).
[21] Scott, D. W., *Virology*, **26**, 85 (1965).
[22] Strauss, J. H., and Sinsheimer, R. L., *J. Mol. Biol.*, **7**, 43 (1963).
[23] Gesteland, R. F., and Boedtker, H., *J. Mol. Biol.*, **8**, 496 (1964).
[24] Mitra, S., Enger, M. D., and Kaesberg, P., *Proc. US Nat. Acad. Sci.*, **50**, 68 (1963).
[25] Boedtker, H., *Biochemistry*, **6**, 2718 (1967).
[26] Isenberg, H., Cotter, R. I., and Gratzer, W. B., *Biochim. Biophys. Acta*, **232**, 184 (1971).
[27] McPhie, P., Hounsell, J., and Gratzer, W. B., *Biochemistry*, **5**, 988 (1966).
[28] Gould, H. J., *Biochemistry*, **5**, 1103 (1966).
[29] Adams, A., Lindahl, T., and Fresco, J. R., *Proc. US Nat. Acad. Sci.*, **57**, 1684 (1967).
[30] Strauss, J. H., and Sinsheimer, R. L., *J. Mol. Biol.*, **34**, 453 (1968).
[31] Fukami, H., and Imahori, K., *Proc. US Nat. Acad. Sci.*, **68**, 570 (1971).
[32] Richards, E. G., Coll, J. A., and Gratzer, W. B., *Analyt. Biochem.*, **12**, 452 (1965).
[33] Bishop, D. H. L., Claybrook, J. R., and Spiegelman, S., *J. Mol. Biol.*, **26**, 373 (1967).
[34] Tinoco, I., Uhlenbeck, O. C., and Levine, M. D., *Nature*, **230**, 362 (1971).
[35] Cory, S., Spahr, P. F., and Adams, J. M., *Cold Spring Harbor Symp. Quant. Biol.*, **35**, 1 (1970).
[36] Gupta, S. L., Chen, J., Schaefer, L., Lengyel, P., and Weissman, S. M., *Biochem. Biophys. Res. Commun.*, **39**, 883 (1970).
[37] Lodish, H. F., and Zinder, N. D., *J. Mol. Biol.*, **19**, 333 (1966).
[38] Gussin, G. N., *J. Mol. Biol.*, **21**, 435 (1966).
[39] Engelhardt, D. L., Webster, R. E., and Zinder, N. D., *J. Mol. Biol.*, **29**, 45 (1967).
[40] Roberts, J. W., and Gussin, G. N., *J. Mol. Biol.*, **30**, 565 (1967).
[41] Zinder, N. D., and Cooper, S., *Virology*, **23**, 152 (1964).
[42] Webster, R. E., Engelhardt, D. L., Zinder, N. D., and Konigsberg, W., *J. Mol. Biol.*, **29**, 27 (1967).
[43] Tooze, J., and Weber, K., *J. Mol. Biol.*, **28**, 311 (1967).
[44] Van Montagu, M., *Arch. Intern. Physiol. Biochim.*, **74**, 941 (1966).
[45] Fiers, W., Van Montagu, M., De Wachter, R., Haegeman, G., Min Jou, W., Messens, E., Remaut, E., Vandenberghe, A., and Van Styvendaele, B., *Cold Spring Harbor Symp. Quant. Biol.*, **34**, 697 (1969).

Printed in Great Britain by Flarepath Printers Ltd., St. Albans, Herts.