Molecular cloning and nucleotide sequence of the 30K and the coat protein cistron of TMV (tomato strain) genome

Nobuhiko Takamatsu, Takeshi Ohno, Tetsuo Meshi and Yoshimi Okada

Department of Biophysics and Biochemistry, Faculty of Science, University of Tokyo, Hongo, Tokyo 113, Japan

Received 11 April 1983; Accepted 9 May 1983

ABSTRACT

The cDNA copies of tobacco mosaic virus (TMV)-tomato strain (L) genome were cloned by the method of Okayama and Berg (Mol. Cell. Biol. 2, 161-170.(1982)) and the sequence of 1,614 nucleotides at the 3' end was determined. The sequence encompasses the 30K and the coat protein cistron which are located in residues 685-1,479 and 203-682 from the 3' end of the genome respectively. The close relationship between the tomato and the common strain was shown on the level of the nucleotide sequence. Highly homologous regions are found in the 3' non-coding region, the assembly origin and the 5' flanking region of the 30K protein cistron. The comparison of the deduced amino acid sequence between the tomato and the common strain shows that the 30K protein is composed of the conserved N-terminal four-fifth and the highly divergent region near the C-terminus.

INTRODUCTION

The RNA genome of tobacco mosaic virus (TMV), a member of tobamovirus group, is about 6,400 nucleotides long and carries the information at least four polypeptides (1, 2). The complete nucleotide sequence of the genomic RNA for a common strain of TMV, vulgare, has been recently determined and reveals that the genomic RNA can encode 130K, 180K, 30K and coat protein (3). The structure and characteristics of the coat protein have been extensively studied and viruses in tobamovirus group have been classified into 10 definitive members on the basis of the amino acid composition of their coat proteins (4). The tomato strain of TMV has been proposed to be most closely related to the common strain of TMV among the viruses examined. The nucleotide sequence homology between the two genomes is also detected by hybridization analysis (5).

TMV particles are reconstituted from its RNA genome and coat protein *in vitro*. The reaction initiates by binding of a specific region (assembly origin) on the RNA to a 20S coat protein aggregate (6). On the basis of the location of the assembly origin, viruses of the tobamovirus group are divided into two subgroups (7). The tomato strain together with the common strain belongs to subgroup 1 in which assembly origin is 800-1,000 nucleotides away from the 3' end of the genomic RNA (8, 9, 10). The cowpea strain (Cc) and cucumber green mottle mosaic virus (CGMMV), belonging to subgroup 2, have their assembly origins at 300-500 nucleotides away from the 3' end (7, 10). The comparison of the genome structure of viruses within a subgroup and between subgroups is interesting from the point of view of virus evolution.

TMV-L is a Japanese tomato strain, isolated from tomato in Japan in 1961 (11). Several mutants have been isolated from the L strain. $L_{11}A$ is an attenuated mutant (12) and used to protect tomato plants against the infection with severe strain in the fields. Ls1 is also an L-derived mutant, which is temperature-sensitive in cell-to-cell movement of virus (13). It was speculated that the 30K protein might be involved in cell-to-cell movement of virus from the comparison of the peptide maps of the 30K proteins of the L and the Ls1 strain (14). The analyses of these mutants and their comparison with the parent strain L will give us the keys to elucidate the functions of non-structural proteins which will relate to virus transport and pathogenecity of TMV.

In this paper, we report the sequence of 1,614 nucleotides at the 3' end of the L-genomic RNA encompassing the 30K and the coat protein cistron, and show the close relationship between the tomato and the common strain on the level of the nucleotide sequence.

MATERIALS AND METHODS

Enzymes

Poly A polymerase was purified from *E. coli* B/r (15). Reverse transcriptase was kindly supplied by Dr. A. Ishihama, Kyoto University. *E. coli* DNA polymerase I, *E. coli* RNase H and terminal deoxynucleotidyl transferase were purchased from BRL and *E. coli* DNA ligase and T4 RNA ligase from P-L Biochemicals. Restriction enzymes, polynucleotide kinase and bacterial alkaline phosphatase were obtained from Takara Shuzo Co.

Preparation of virus and viral RNA

TMV-L particles were purified from infected *Nicotiana tabacum* L. cv. Xanthi (16). RNA was extracted from purified virus with phenol and SDS (17).

Construction of cDNA clones

TMV-L RNA was polyadenylated at the 3' end as described in our previous paper (17), followed by Sephadex G-100 gel filtration to remove ATP. The cDNA copies were cloned by the method of Okayama and Berg (18) with some modifications. The vector-primer was prepared from the plasmid p3-2-1, the pBR322-SV40 recombinant, containing a SV40 DNA segment between the PvuII and HindIII sites of pBR322 DNA (19). The linker DNA, a SV40 DNA segment having oligo dG-tail at one end and a HindIII cohesive end at the other, was prepared from pX as described (18). p3-2-1 and pX were kindly supplied by Dr. K. Oda at University of Tokyo. cDNA was synthesized at 42°C for 60 min under the condition that was 50 mM Tris-HCl (pH7.9 at 42°C), 12.5 mM $MgCl_2$, 30 mM KCl, 10 mM DTT, 4 mM NaPPi, 1 mM each dATP, dTTP, dGTP and [^{3}H]-dCTP (230 cpm/pmol), 50μg/ml polyadenylated L-genomic RNA, 25μg/ml the vector-primer and 100 units/ml reverse transcriptase. In the ligation step, 300 units/ml E. coli DNA ligase was used. The nick-translation reaction was carried out using 45 units/ml E. coli DNA polymerase I, 200 units/ml E. coli DNA ligase and 4 units/ml E. coli RNase H to replace RNA template by DNA. The reaction mixture was incubated at 12°C for 1 hr and at 30°C for another 3 hr.

Screening of recombinant clones

Transformation was carried out as described (18), and ampicillin-resistant transformants were screened by the colony hybridization according to the method of Grunstein and Hogness (20). For probe, [^{32}P]-labeled, fragmented TMV-L RNA was prepared (17).

Nucleotide sequencing

Plasmid DNA was purified by the method of Katz et al. (21) with slight modification. Restriction fragments were prepared using restriction enzymes AvaII, HaeIII, HincII, Sau3AI for pL-1-13 and AccI, BstNI, FokI, HapII, HinfI, PvuII for pL-2-28. DNA sequencing was performed by the method of Maxam and Gilbert (22) with slight modification (23).

The genomic RNA was labeled at the 3' end using T4 RNA ligase and [^{32}P]-pCp (24). RNA sequencing was carried out by the method of Peattie (25).

RESULTS

Cloning of cDNA and the sequence of 1,614 nucleotides at the 3' end of the genomic RNA

For cloning the cDNA copies of TMV-L genomic RNA we used the method described by Okayama and Berg (18). First-strand cDNA synthesis was primed by dT-tail of the vector-primer annealed with poly A-tail attached to the 3' end of the genomic RNA. To generate a cohesive tail at the 3' end of the cDNA, about 20 dC residues were added to the 3' ends of both the cDNA and the

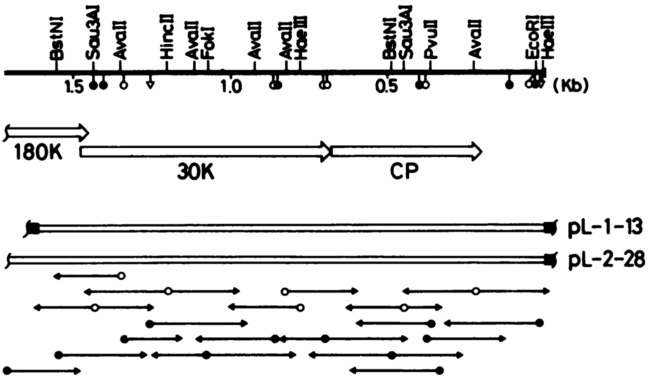

Fig. 1 Restriction map of an about 1,600 nucleotide sequence of L-RNA from the 3' end and strategy for sequencing. Abbrevations are ∇ for HapII, ○ for HinfI and ● for TaqI . The direction and extent of sequence determination are indicated by the arrows. Terminal circles of the arrows indicate the labeled 5' ends of sequenced fragments (○ for pL-1-13 and ● for pL-2-28). The locations of the coat protein (CP), the 30K protein (30K) and the 180K protein cistron (180K) are also shown.

vector DNA and the oligo dC-tail at the vector terminus was removed by cleavage at the unique HindIII site of the vector-primer. The vector-cDNA:RNA derivatives with a HindIII cohesive end and oligo dC-tail at the other were cyclized by E. coli DNA ligase, mediated by the linker DNA. Before transformation, RNA was replaced by second-strand cDNA by nick-translation using E. coli DNA polymerase I and E. coli RNase H.

From 80 ampicillin-resistant transformants, two clones, pL-1-13 and pL-2-28, were selected by colony hybridization, rapid screening procedure (24) and comparison of the restriction maps. pL-1-13 and pL-2-28 carry approximately 1,600 and 2,300 bp cDNA insert respectively. The restriction maps show that pL-2-28 encompasses pL-1-13 (Fig.1).

The nucleotide sequence was first determined with pL-1-13 and in order to confirm it the opposite strand of pL-2-28 was sequenced (Fig.1). The sequence of the 3' end was complemented by sequencing RNA directly and the two clones were proved to carry the nucleotide sequence of the 3' end of the genomic RNA (data not shown). The sequence of 1,614 nucleotides at the 3' end of the L-genomic RNA is presented in Fig.2, together with the sequences of the two members of the TMV common strain (vulgare and OM) (3, 26 and unpublished data). In the region sequenced there is no mismatch between the two L cDNA clones except a difference in length of A cluster in residues 43–49 from the 3' end of the genomic RNA. Four clones were further selected to determine the sequence of the region. The result was that four clones out of six had seven A residues and two had eight. The banding pattern of RNA sequencing gel also showed the heterogeneous length of the A cluster, ranging from 6 residues to 9 (data not shown). Consequently, the heterogeneous length of the A cluster does not result from the artifact in the process of the cloning but from the polymorphism of the genomic RNA.

Coding regions for 180K, 30K and coat protein

The coat protein cistron of TMV-L is located in residues 203-682 from the 3' end of the genomic RNA, starting from AUG at residues 680-682 and terminated by UAA at residues 203-205. The 3' non-coding region is 202 nucleotides long. The coat protein is composed of 158 amino acids as that of the common strain and its calculated molecular weight is 17,641, considering N-acetylation (Nishiguchi, personal communication). The amino acid sequence of the coat protein of another tomato strain, dahlemense, has been reported (27). Between the coat proteins of the two tomato strains, two substitutions were found: Asn at amino acid position 29 and Ala at 86 in the L strain are replaced by Ser and Thr in the dahlemense strain respectively. From the alignment of the nucleotide sequence of the L strain with that of the common strain the capping site of the coat protein mRNA of the L strain would be predicted to be a G residue at 691 (Fig.2).

The 30K protein cistron is located in residues 685-1,479 from the 3' end, starting from AUG at residues 1,477-1,479 and terminated by UAA at residues 685-687, and separated from the following coat protein cistron by two nucleotides. The 30K protein is composed of 263 amino acids, four residues shorter than that of the common strain (3, 26), and its calculated molecular weight is 29,186. The protein is characterized by its high content of charged amino acids (36 acidic amino acids and 39 basic amino acids).

The complete nucleotide sequence of the genomic RNA of the common strain reveals that the 180K protein cistron overlaps by five codons with the 30K protein cistron (3). From the alignment of the nucleotide sequences of the L and the common strain the predicted 180K protein cistron of the L strain probably terminates by UAA at residues 1,463-1,465 in the 30K protein cistron (Fig.2).

The structure of assembly origin

The assembly origin of the tomato strain is located in the same region on the genomic RNA as that of the common strain (10). The region of the common strain can be folded into a stable hairpin loop structure (9, 28). For the tomato strain a stable hairpin loop structure can also

```
1614 ACAAUUGCGCGUAUUACACACAAUUGGACGACGCUGUUGGGGAGGUUCAUAAAACCGCCCCACCUGGUUCGUUUGUUUAUAAGAGUUUAG L
     -------U--------------G--------------AU-------------G--------U--A-----------------A---C-G- Vulgare
     -------U--------------G--------------AU-------------G--------U--A-----------------A---C-G- OM
                                                  ┌──30K protein cistron ───→
                                                  |  AlaLeuValValLysGlyLysValAsnIleAsnGluPheIle
1524 UUAAGUAUUUGUCAGAUAAAGUUUUGUUUAGAAGUUUAUUUCUUGAUGGCUCUAGUUGUUAAAGGUAAGGUAAAUAUUAAUGAGUUUAUC L
     -G-----------U---------C-U-----------G---A-A---------------------A--A--G-----C------------ Vulgare
     -G-----------U---------C-U-----------G---A-A---------------------A--A--G-----C------------ OM
                             ←────( 180K protein cistron )────────┘
     AspLeuSerLysSerGluLysLeuLeuProSerMetPheThrProValLysSerValMetValSerLysValAspLysIleMetValHis
1434 GAUCUGUCAAAGUCUGAGAAACUUCUCCCGUCGAUGUUCACGCCUGUAAAGAGUGUUAUGGUUUCAAAGGUUGAUAAGAUUAUGGUCCAU L
     --C---A----AAUG-----GA-CU-A-----------U--C--------------U---UG---C--A--------A--A-----U--- Vulgare
     --C---A----AAUG-----GA-CU-A-----------U--C--------------C---UG---C--A--------A--A-----U--- OM
           Thr   Met      Ile                                    Cys
                     50
     GluAsnGluSerLeuSerGluValAsnLeuLeuLysGlyValLysLeuIleGluGlyGlyTyrValCysLeuValGlyLeuValValSer
1344 GAAAAUGAAUCAUUGUCUGAAGUAAAUCUCUUAAAAGGUGUAAAACUUAUAGAAGGUGGGUAUGUUUGCUUAGUCGGUCUUGUUGUGUCC L
     --G-----G--------A--G--G--C--UC-U-----A--U--G-----U--UA----A--C--C--U----C----U-G--C--CA-G Vulgare
     --G-----G--------A--G-----C--UC-C-----A--U--G-----U--UA----A--C--C--U----C----U-G--C--CA-G OM
                                                        AspSer               Ala            Thr
                                                                                100
     GlyGluTrpAsnLeuProAspAsnCysArgGlyGlyValSerValCysMetValAspLysArgMetGluArgAlaAspGluAlaThrLeu
1254 GGUGAGUGGAAUUUACCAGAUAAUUGCCGUGGUGGUGUGAGUGUCUGCAUGGUUGACAAGAGAAUGGAAAGAGCGGACGAAGCCACACUG L
     --C--------C--G--U--C------A-A--A--------C--G--UC----G-----A--G-----------C-----G-----U--C Vulgare
     --C--------C--G--U--C------A-A--A--------C--G--UC----G-----A--G-----------C-----G-----U--C OM
                                                    Leu

     GlySerTyrTyrThrAlaAlaAlaLysLysArgPheGlnPheLysValValProAsnTyrGlyIleThrThrLysAspAlaGluLysAsn
1164 GGGUCAUAUUACACUGCUGCUGCUAAAAAGCGGUUUCAGUUUAAAGUGGUCCCAAAUUACGGUAUUACAACAAAGGAUGCAGAAAAGAAC L
     --A--U--C-----A--A-----A--G--AA-A--------C--G--C--U--C-----U-C---A--C--CC----C--GAUG--A--- Vulgare
     --A--U--C-----A--A-----A--G--AA-A--------C--G--C--U--C-----U-C---A--C--CC----C--GAUG--A--- OM
                                                                 Ala         Gln      Met
                                                  150
     IleTrpGlnValLeuValAsnIleLysAsnValLysMetSerAlaGlyTyrCysProLeuSerLeuGluPheValSerValCysIleVal
1074 AUAUGGCAGGUCUUAGUAAAUAUUAAAAAUGUAAAAAUGAGUGCGGGCUACUGCCCUUUGUCAUUAGAAUUUGUGUCUGUGUGUAUUGUU L
     G-C-----A--U-----U-------G------G--G---UCA-----U-U---U--GC-U--UC-G--G--------G------------ Vulgare
     G-G-----A--U-----C-------G---------G---UCA-----U-U---U--GC-U--UC-G--G--------G--------C--- OM
     Val                     Arg                     Phe

     TyrLysAsnAsnIleLysLeuGlyLeuArgGluLysValThrSerValAsnAspGlyGlyProMetGluLeuSerGluGluValValAsp
 984 UAUAAAAAUAAUAUAAAAUUGGGUUUGAGGGAGAAAGUAACGAGUGUGAACGAUGGAGGACCCAUGGAACUUUCGGAAGAAGUUGUUGAU L
     ----G---------------A--------A-----GA-U--A-AC----GA--C-----G------------A-A--------C------ Vulgare
     ----G---------------A--------A-----GA-C--A-------GA--------G------------A-A--------------- OM
        Arg                              Ile         Arg                     Thr
                    200
     GluPheMetGluAsnValProMetSerValArgLeuAlaLysPheArgThrLysSerSerLysArgGlyProLysAsnAsnAsnAsnLeu
 894 GAGUUCAUGGAGAAUGUUCCAAUGUCGGUUAGACUCGCAAAGUUUCGAACCAAAUCCUCAAAAAGAGGUCCGAAAAAUAAUAAUAAUUUA L
     -----------AG----C--U------A-C--G--U------------         --UCG--CC--AAAA--G-G-G--GUCCGCAA- Vulgare
     -----------AG----C--U-----AA-C--G--U------------         --UCG--CC--AAAA--G-G-G--GUCCG-AA- OM
                 Asp            Ile                              ArgThr   Lys   SerAspValArgLys

     GlyLysGlyArgSerGlyGlyArgProLysProLysSerPheAspGluValGluLysGluPheAsp                     Asn
 804 GGUAAGGGGCGUUCAGGCGGAAGGCCUAAACCAAAAAGUUUUGAUGAAGUUGAAAAAGAGUUUGAU                     AAU L
     --G--AAAUA--AGUAAU-AUC--U-AGUG--G--C-AGAACU--AG-AA--UU--G--U----GAGGAAUGAGUUUUAAAAAGAAU--- Vulgare
     --G--AAUUA--AGUA-U-AUC--U-AGCG--G--C-AGAACU--AG-AA--UU--G--U----GAGGAAUGAGUUUUAAAAAGAAU--- OM
           IleSer   SerAsp   SerAla   AsnLysAsnTyrArgAsnVal   Asp   GlyGlyMetSerPheLysLysAsn
           250                                           ┌──────Coat protein cistron ────→
     LeuIleGluAspGluAlaGluThrSerValAlaAspSerAspSerTyr     |  SerTyrSerIleThrSerProSerGlnPheVal
 735 UUGAUUGAAGAUGAAGCCGAGACGUCGGUCGCGGAUUCUGAUUCGUAUUAAAUAUGUCUUACUCAAUCACUUCUCCAUCGCAAUUUGUG L
     --A--C--U-----UU-G---G-UA-U-----C--A--G-------U---------------AGU------A-------U--G--C--- Vulgare
     --A--C--U-----UU-G-----UA-U-----C--A--G-------U---------------AGU------A-------U--G--C--- OM
           Asp   AspSer      Thr      Glu        ↑Phe                       Thr
```

```
    PheLeuSerSerValTrpAlaAspProIleGluLeuLeuAsnValCysThrAsnSerLeuGlyAsnGlnPheGlnThrGlnGlnAlaArg
646 UUUUUGUCAUCUGUAUGGGCUGACCCUAUAGAAUUGUUAAACGUUUGUACAAAUUCGUUAGGUAACCAGUUUCAAACACAGCAAGCAAGA L
    --C--------A-CG-----C-----A-----G--AA-U--UU-A-----U---G-C-----A--U--------------A-----UC-- Vulgare
    --C------G-A-CG-----C-----A-----G--AA-U--UU-A-----U---G-C-----A--U--------------A-----UC-- OM
             AlaAla                     Ile   Leu         Ala

                             50
    ThrThrValGlnGlnGlnPheSerGluValTrpLysProPheProGlnSerThrValArgPheProGlyAspValTyrLysValTyrArg
556 ACUACUGUUCAACAGCAGUUCAGCGAGGUGUGGAAACCUUUCCCUCAGAGCACCGUCAGAUUUCCUGGCGAUGUUUAUAAGGUGUACAGG L
    ---GUC------AGA--A-----U----------------CA--A--AGUA--U--U--G--C----A-AG--AC-U------------- Vulgare
    ---GUC------AGA--A-----U----------------CA--A--AGUG--U--U--G--C----A-AG--AC-U------------- OM
       Val      Arg                        Ser      Val               AspSerAspPhe

                                                                                        100
    TyrAsnAlaValLeuAspProLeuIleThrAlaLeuLeuGlyAlaPheAspThrArgAsnArgIleIleGluValGluAsnGlnGlnSer
466 UACAAUGCAGUUUUAGAUCCUCUAAUUACUGCGUUGCUGGGGGCUUUUGAUACUAGGAAUAGAAUAAUCGAAGUAGAAAACCAGCAGAGU L
    --------G--A-----C--G---G-C--A--AC--U-A--U--A--C--C-----A-----------A-----U-----U---GC--AC Vulgare
    --------G--A-----C--G---G-C--A--AC--U-A--U--A-----C-----A-----------A-----U-----U---GC--AC OM
                            Val                                                         AlaAsn

    ProThrThrAlaGluThrLeuAspAlaThrArgArgValAspAspAlaThrValAlaIleArgSerAlaIleAsnAsnLeuValAsnGlu
376 CCGACAACAGCUGAAACGUUAGAUGCUACCCGCAGGGUAGACGACGCUACGGUUGCAAUUCGGUCUGCUAUAAAUAAUUUAGUUAAUGAA L
    --C--G--U--C-----------------U--U--A-----------A-----G--C--AA--AGC--G------------A-AGUA--- Vulgare
    --C-----U--C-----------------U--U--A-----------A-----G--C--AA--AGC-----------------AGUA--- OM
                                                                                        Val

                                                           150
    LeuValArgGlyThrGlyLeuTyrAsnGlnAsnThrPheGluSerMetSerGlyLeuValTrpThrSerAlaProAlaSer
286 CUAGUAAGAGGUACUGGACUGUACAAUCAGAAUACUUUUGAAAGUAUGUCUGGGUUGGUCUGGACCUCUGCACCUGCAUCUUAA  AUGC L
    U-GA-C-----A--C---UCU--U----G--GCU----C--G--CUCU-----U-----U----------GU------A---G-GGUA-U Vulgare
    U-GA-C-----A--C---UCU--U----G--GCU----C--G--CUCU-----U-----U----A-----GU------A---G-GGUA-U OM
       Ile            Ser      ArgSerSer         Ser               Asn   Gly      Thr

198 AUAGGUGC   UGAAAUAUAAAGUUUGUGUUUCUAAAACACACGUGGUACGUACGAUAACGUACAGUGUUUUUCCCUCCACUUAAAUCGA L
    CA--A---AUAAU----- -CG-A------CCG- --U----------G------------C-U-------------------------- Vulgare
    CA--A---AUAAU----- -CG-A------CCG- --U----------G------------C-U-------------------------- OM

                                                                     ⇩
111 AGGGUAGUGUCUUGGAGCGCGCGGAGUAAACAUAUAUGGUUCAUAUAUGUCCGUAGGCACGUAAAAAAAGCGAGGGAUUCGAAUUCCCCC L
    -----U----------U-------GUC---UG---------------CA----C-------- --U----------G-------C----- Vulgare
    -----U----------U-------GUC---UG---------------CA----C-------- --U----------G-------C----- OM

 21 GGAACCCCCGGUUGGGGCCCAOH
    -UU---------A--------
    -UU---------A--------
```

Fig. 2 Sequence of 1,614 nucleotides at the 3' end of TMV L-genomic RNA and deduced amino acid sequences of the coat and the 30K protein. The nucleotide sequences of OM- and vulgare-RNA are aligned (3, 26 and unpublished data). Hyphens denote nucleotides identical to those in the L sequence. The absence of a hyphen or a base indicates deletion. Only amino acids different from those of the L proteins are denoted. The capping site of the coat protein mRNA for a common strain (⬆)(8) and the A cluster showing heterogeneity (⇩) are indicated. The nucleotide is numbered from the 3' end of the genomic RNA, regarding the length of the A cluster as seven residues long that is the major length.

be constructed in residues 850-950 (Fig.3). A few base substitutions found between the L and the common strain are not proved to destroy the hairpin loop structure. The common features found in the assembly origin of

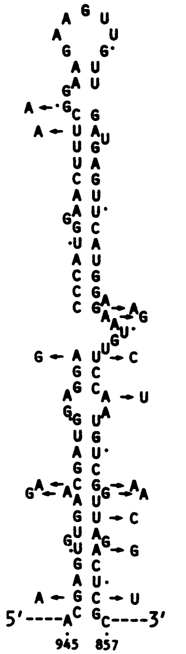

Fig. 3 A possible secondary structure folded in the assembly origin (residues 857-945). The thermodynamic stability was calculated to be -25.2 kcal/mol following the rules of Tinoco et al. (29). Different nucleotides of OM-RNA from those of L-RNA are indicated.

other strains, vulgare, OM, Cc and CGMMV, triplet-repeated purine base tract and postulated target sequence (-GAPuGUUG-) at the loop of the hairpin structure (24, 30), are found in that of the L strain.

DISCUSSION

The tomato strain has been considered to be the most closely related to the common strain on the basis of the amino acid composition of the coat protein (4). Our data shows that the close relationship is shown on the level of the nucleotide sequence. About 75 % of nucleotides match between the tomato and the common strain from the alignment shown in Fig.2.

Fig.4 shows the magnitude of the nucleotide sequence homology between the L and the OM strain averaged for 100 nucleotides. Highly homologous regions with the homology above 85% are found in residues 1-170, 850-1,000 and 1,460-1,614. The first region is the 3' non-coding region and the homology would be due to the conservative pressure of the viral replication mechanism. The second corresponds to the assembly origin. The third is the 5' flanking region of the 30K protein cistron. The capping site of the subgenomic mRNA for the 30K protein will be in this region, but the highly homologous region extends over much longer sequence than the 5' flanking

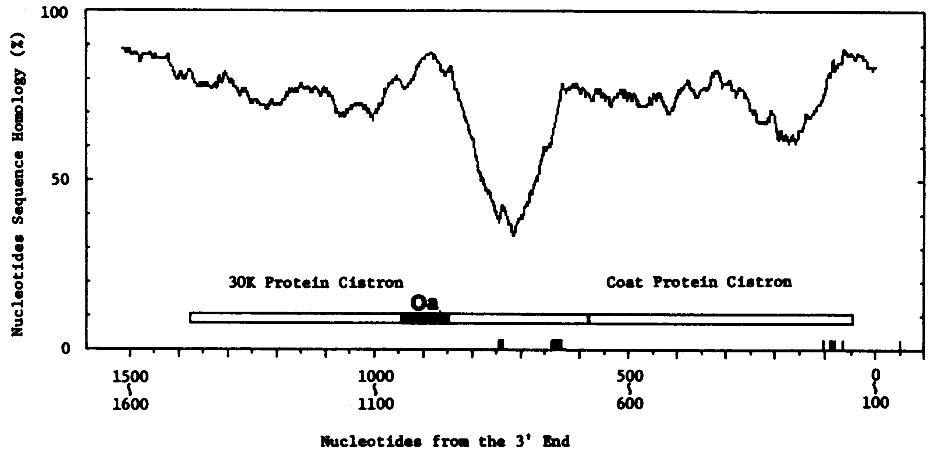

Fig. 4 Nucleotide sequence homology between TMV L and OM strain. The homology is shown as the percentage of the common nucleotides in successive 100 nucleotides. The gaps inserted to maximize the homology as in Fig.2 are indicated by the short vertical lines on the transversal axis. The locations of the cistrons and the assembly origin (Oa) are also shown.

region of the coat protein cistron. This region will code the C-terminal portion of the 180K protein, but the homology is rather high compared with other coding regions, considering the degeneracy of the genetic codes. The nucleotide sequence itself may bear some function as the other two homologous sequences.

On the other hand, highly divergent nucleotide sequences are found in regions coding for the C-termini of the coat and the 30K protein, and their amino acid sequences are also fairly different. Comparing the amino acid sequences of the coat proteins of the L and the OM strain, 27 replacements are found in total 158 amino acids. Most of them occur in both the N- and the C-terminal portion, especially in residues 130-158 at the C-terminus, where 10 residues out of 28 are replaced.

Comparing the L-30K protein with that of the OM strain, the protein can be divided into two subregions on the basis of the extent of the amino acid sequence homology (Fig.5). The regions corresponding to residues 1-210 of L-30K protein, four-fifth of the protein, are highly homologous (about 90%). In the contrast, the homology reduces drastically in the C-terminal portion, especially in residues 211-246, where the homology is only 30%. In the latter region the amino acid compositions are still similar and basic amino acids are abundant (10 out of 36 in L and 12 out of 40 in OM). The

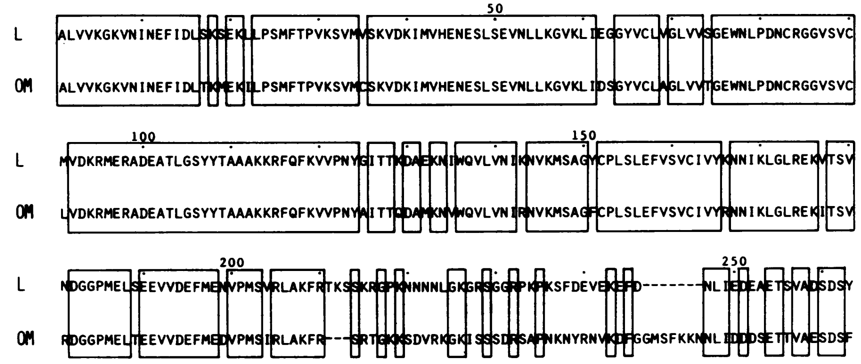
50
L ALVVKGKVNINEFIDLSKSEKLLPSMFTPVKSVMVSKVDKIMVHENESLSEVNLLKGVKLIEGGYVCLVGLVVSGEWNLPDNCRGGVSVC
OM ALVVKGKVNINEFIDLTKMEKILPSMFTPVKSVMCSKVDKIMVHENESLSEVNLLKGVKLIDSGYVCLAGLVVTGEWNLPDNCRGGVSVC
100
150
L MVDKRMERADEATLGSYYTAAAKKRFQFKVVPNYGITTKDAEKNIWQVLVNIKNVKMSAGYCPLSLEFVSVCIVYKNNIKLGLREKVTSV
OM LVDKRMERADEATLGSYYTAAAKKRFQFKVVPNYAITTQDAMKNVWQVLVNIRNVKMSAGFCPLSLEFVSVCIVYRNNIKLGLREKITSV
200
250
L NDGGPMELSEEVVDEFMENVPMSVRLAKFRTKSSKRGPKNNNNLGKGRSGGRPKPKSFDEVEKEFD-------NLIEDEAETSVADSDSY
OM RDGGPMELTEEVVDEFMEDVPMSIRLAKFR---SRTGKKSDVRKGKISSSDRSAPNKNYRNVKDFGGMSFKKNNLIDDDSETTVAESDSF

Fig. 5 Comparison of the amino acid sequence of the 30K protein of the L strain with that of the OM strain. Common residues are boxed.

30K protein is a most probable candidate involved in the cell-to-cell movement of virus (14), and the host range of a plant virus has been speculated to be dependent upon whether a virus can spread into other cells from the initially inoculated cell (31). The similarity of the 30K proteins of the L and the OM strain is not incompatible with overlap of the host range. The conserved N-terminal four-fifth region might have fundamental function in virus transport and the highly divergent region near the C-terminus may reflect minor difference of the symptoms on some plants (11).

We have classified several viruses belonging to tobamovirus group into subgroup 1 (common and tomato strain) and subgroup 2 (cowpea strain and CGMMV) on the basis of the location of the assembly origin (7). The sequences of 1,000-2,000 nucleotides from the 3' end of the genome of all four strains were determined (3, 24, 26, 30, 32) and their genome structures are summarized in Fig.6. The four strains are classified into two groups from their genome structures and this grouping coincides with that from the location of the assembly origin. The genome structures of two members of the common strain, vulgare and OM, are identical. The coat and the 30K protein cistron are located in residues 205-684 and 687-1,493 from the 3' end of the genomic RNA respectively, and the two cistrons are separated by two nucleotides. The 180K protein cistron overlaps the 30K protein cistron (3, 26). The genome structure of the tomato strain is very similar to that of the common strain. There is a two base-gap between the coat and the 30K protein cistron and the presumed 180K protein cistron overlaps the 30K protein cistron. On the other hand, the nucleotide sequence

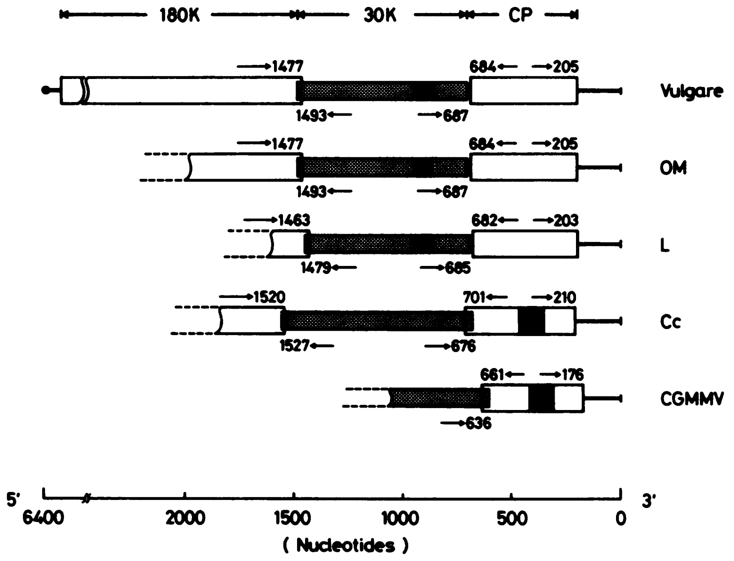

Fig. 6 Diagramatic summary of genome structure at the 3' end portion of several tobamoviruses. Coding regions are indicated by boxes. The number shows the first letter of the initiation codon or the third letter of the termination codon from the 3' end of the genomic RNA. Darkened area shows the assembly origin.

of the cowpea strain shows overlap of the coat and the 30K protein cistron, and the 30K protein cistron is deduced to overlap an open frame (probably of the 180K protein cistron) at the other end (32). The overlap of the coat and the 30K protein cistron is also found in CGMMV genome (24). Although the meaning of the coincidence of the grouping from the location of the assembly origin and that from the genome structure is not known for the present, this grouping is also consistent with the evolutionary tree deduced from the amino acid compositions or sequences of the coat proteins of tobamoviruses by Gibbs (33). As the common and the tomato strain, and the cowpea strain and CGMMV are most distantly related, to elucidate the evolution of the tobamovirus, further analysis of the site of the assembly origin and genome structure of other intermediate strains will be necessary and interesting.

ACKNOWLEDGMENTS

We thank Drs. M. Nishiguchi and Y. Kiho at Institute for Plant Virus Research for preparation of TMV-L, Dr. K. Oda at University of Tokyo for plasmids p3-2-1 and pX and Dr. A. Ishihama at Kyoto University for reverse transcriptase. This work was supported in part by a Grant-in-Aid for Special Project Research from the Ministry of Education, Science and Culture, Japan.

REFERENCES

1. Beachy, R.N., Zaitlin, M., Bruening, G. and Israel, H.W. (1976). Virology **73**, 498-507.
2. Hunter, T.R., Hunt, T., Knowland, J. and Zimmern, D. (1976). Nature (London) **260**, 759-764.
3. Goelet, P., Lomonossoff, G.P., Butler, P.J.G., Akam, M.E., Gait, M.J. and Karn, J. (1982). Proc. Natl. Acad. Sci. USA **79**, 5818-5822.
4. Gibbs, A.J. (1977). CMI/AAB Descrip. Plant Viruses, No.184.
5. Van De Walle, M.J. and Siegel, A. (1982). Phytopathology **72**, 390-395.
6. Hirth, L. and Richards, K.E. (1981). Adv. Virus Res. **26**, 145-199.
7. Fukuda, M., Meshi, T., Okada, Y., Otsuki, Y. and Takebe, I. (1981). Proc. Natl. Acad. Sci. USA **78**, 4231-4235.
8. Guilley, H., Jonard, G., Kukla, B. and Richards, K.E. (1979). Nucl. Acids Res. **6**, 1287-1308.
9. Zimmern, D. (1977). Cell **11**, 463-482.
10. Fukuda, M., Okada, Y., Otsuki, Y. and Takebe, I. (1980). Virology **101**, 493-502.
11. Oshima, N., Goto, T. and Sato, R. (1964). Research Bulletin No.83 of the Hokkaido National Agricultural Experimental Station, pp.87-99.
12. Goto, T. and Nemoto, M. (1971). Research Bulletin No.99 of the Hokkaido National Agricultural Experimental Station, pp.67-76.
13. Nishiguchi, M., Motoyoshi, F. and Oshima, N. (1978). J. gen. Virol. **39**, 53-61.
14. Leonard, D.A. and Zaitlin, M. (1982). Virology **117**, 416-424.
15. Sippel, A.E. (1973). Eur. J. Biochem. **37**, 31-40.
16. Motoyoshi, F. and Oshima, N. (1975). J. gen. Virol. **29**, 81-91.
17. Meshi, T., Takamatsu, N., Ohno, T. and Okada, Y. (1982). Virology **118**, 64-75.
18. Okayama, H. and Berg, P. (1982). Mol. Cell. Biol. **2**, 161-170.
19. Oda, K., Kato, H., Saito, I., Sugano, S., Maruyama, K., Masuda, M., Shiroki, K. and Shimojo, H. (1983). J. Virol. **45**, 408-419.
20. Grunstein, M. and Hogness, D.S. (1975). Proc. Natl. Acad. Sci. USA **72**, 3961-3965.
21. Katz, L., Kingsbury, D.T. and Helinsky, D.R. (1973). J. Bacteriol. **114**, 577-591.
22. Maxam, A.M. and Gilbert, W. (1980). In "Methods in Enzymology" (Grossman, L., ed.) Vol.65, pp.499-560. Academic Press, New York.
23. Bernard, O. and Gough, N.M. (1980). Proc. Natl. Acad. Sci. USA **77**, 3630-3634.
24. Meshi, T., Kiyama, R., Ohno, T. and Okada, Y. Virology in press.
25. Peattie, D.A. (1979). Proc. Natl. Acad. Sci. USA **76**, 1760-1764.
26. Meshi, T., Ohno, T. and Okada, Y. (1982). J. Biochem. **91**, 1441-1444.
27. Wittman-Liebold, B. and Wittman, H.G. (1963). Z. Vererbungs. **94**, 427-435.
28. Jonard, G., Richards, K.E, Guilley, H. and Hirth, L. (1977). Cell **11**, 483-493.
29. Tinoco, J.I., Borer, P.N., Dengler, B., Levin, M.D., Uhlenbeck,O.C., Crothers, D.M. and Gralla, S. (1973). Nature New Biol. **246**, 40-41.
30. Meshi, T., Ohno, T., Iba, H. and Okada, Y. (1981). Mol. Gen. Genet. **184**, 20-25.
31. Taliansky, M.E., Malishenko, S.I., Pshennikova, E.S and Atabekov, J.G. (1982). Virology **122**, 327-331.
32. Meshi, T., Ohno, T. and Okada, Y. (1982). Nucl. Acids Res. **10**, 6111-6117.
33. Gibbs, A. (1980). Intervirology **14**, 101-108.