Sequence analysis of rice hoja blanca virus RNA 3

Joachim de Miranda, Myriam Hernandez, Roger Hull and Ana M. Espinoza

Centro de Investigación en Biología Celular y Molecular, Universidad de Costa Rica, Costa Rica and John Innes Institute, Norwich Research Park, Colney Lane, Norwich NR4 7UH, U.K.

RNA 3 of rice hoja blanca tenuivirus (RHBV) has 2299 nucleotides and resembles RNA 3 of other tenuiviruses such as maize stripe (MSIV) and rice stripe (RSIV) viruses in potentially coding with an ambisense strategy for two proteins. Both the viral-sense protein of 23K and the complementary-sense protein of 35K have about 46% amino acid identity with the analogous proteins encoded by RNA 3 of MSIV and RSIV. As the proteins of MSIV and RSIV have about 46% identity between themselves, RHBV cannot be a South and Central American strain of the Asian RSIV. The intergenic region resembles those of other tenuiviruses, being rich in A and U residues, but its predicted folding pattern is unlike those of other tenuiviruses. Instead, the predicted folding of the intergenic region was indistinguishable from that of the coding regions and there was no evidence for a distinct hairpin-loop structure. The significance to the evolution of tenuiviruses of the similarities that the two proteins have with their analogues in other tenuiviruses is discussed.

Rice hoja blanca virus (RHBV) is a member of the tenuivirus group, whose other known members are rice stripe virus (RSIV), maize stripe virus (MSIV), European wheat striate mosaic virus and rice grassy stunt virus. Tenuiviruses have a highly characteristic genome, typically composed of four or five ssRNA species that total 16 to 18 kb (for a review see Ramirez & Haenni, 1994). RHBV has four visible ssRNA species of approximately 9800, 3500, 2300 and 1900 nucleotides and their corresponding double-stranded forms, which are presumed to arise from annealing of separately encapsidated complementary RNAs (Ramirez et al., 1992). Sequence analysis of RNAs 2, 3 and 4 of RSIV (Kakutani et al., 1990, 1991; Zhu et al., 1991, 1992; Takahashi et al., 1993) and RNAs 3 and 4 of MSIV (Huiet et al., 1991, 1992) has shown that each of these RNAs has an ambisense translation strategy. RNA 3 of RSIV and MSIV encodes a non-structural protein (NS3) at the 5' end of the viral-sense RNA and the nucleocapsid protein (N) at the 5' end of the complementary-sense RNA.

For all the ambisense RNAs, the coding regions are separated by a large intergenic region which may be capable of RNA folding. The 5' and 3' termini of each RNA are complementary, forming panhandles, which are characteristic of bunyaviruses and tospoviruses (Elliott, 1990; German et al., 1992). These terminal sequences are highly conserved among all tenuivirus RNAs analysed so far, and the eight terminal nucleotides are the same as those of the S, M and L RNAs of the phlebo- and uukuviruses (Kakutani et al., 1991; Schmaljohn & Patterson, 1990). In this paper we present the sequence of RHBV RNA 3 and compare its features with those of RSIV and MSIV.

Ribonucleoprotein particles were purified as described by Morales & Niessen (1983) from rice infected with a Costa Rican isolate of RHBV. One μg of viral RNA extracted from the ribonucleoprotein was converted to cDNA (Kotewicz et al., 1988) using 2 μm of the oligonucleotide 5' CCCGGGCGGCCACACAAA-GTC Y. The last 10 bases of this primer (bold) correspond to the highly complementary and conserved 5' and 3' termini of all tenuivirus RNAs analysed so far. Consequently the primer could be used for both first- and second-strand cDNA synthesis. The other bases in the oligonucleotide are not complementary to RHBV and contain a NotI restriction site. Twenty-five μl of first-strand cDNA mixture was amplified by PCR in 100 μl of 32.5 mM-Tris-HCl pH 8.0, 3.5 mM-MgCl₂, 44 mM-KCl, 2.5 mM-DTT, 0.1% BSA, 0.3 mM of each dNTP, 1 μm-primer and 5 units of Taq DNA polymerase (these concentrations are inclusive of first-strand reaction components). The PCR thermal profile was two cycles of 92 °C for 1 min, 20 °C for 1 min and 72 °C for 4 min followed by 30 cycles of 92 °C for 1 min, 55 °C for 1 min and 72 °C for 4 min. The PCR products were purified, size-fractionated on gels to select fragments of approximately 500 to 3000 bp, digested with NotI and

The nucleotide sequence data reported in this paper have been submitted to the GenBank and assigned the accession number L07940.
Fig. 1. For legend see opposite.
cloned into pBluescript KS+. Families of related clones were identified by cross-hybridization and by Northern blotting of RHBV genomic RNAs. A selection of clones relating to RNA 3 (2.3 kb) was sequenced using the Sequenase kit (USB) according to the manufacturer's instructions. Since the sequences were obtained from cloned PCR products, there was a greater probability than usual that the sequences contained PCR-induced errors (Reiss et al., 1990). We therefore sequenced both strands of a full-length clone and several overlapping partial clones so that at each nucleotide position the consensus sequence was derived from at least three sequences, with at least two of them being the same. The 5' sequence was confirmed by the method of Fichot & Girard (1990), modified to incorporate a ratio of 1:1 dNTP:ddNTP instead of 2:1. The primer was complementary to nucleotides 53 to 72 of the RHBV sequence. This confirmed the 5' sequence, except for ambiguity with the 5'-terminal A residue. For confirmation of the 3' sequence, virion RNA was polyadenylated using poly(A) polymerase (Gibco BRL) under conditions recommended by the manufacturer and DNA synthesis was primed using oligo-d(T)18. The sequences were analysed using the computer programs of the Genetics Computer Group (Devereux et al., 1984).

RHBV RNA 3 comprises 2299 nucleotides with an overall nucleotide composition of 29% A, 22% C, 18% G and 31% U, which is very similar to that of the RNAs 3 of MStV and RStV. The consensus sequence from several independent clones (Fig. 1) shows that there are two open reading frames (ORFs) which, like the RNAs 3 of MStV and RStV, are in an ambisense translation arrangement. The viral-sense ORF would encode a 23K protein and the complementary sense ORF a 35K protein. They are separated by a large intergenic spacer, rich in A (31%) and especially U (38%) residues. Clones were specifically created to confirm that the full-length clone did indeed represent the full genome, since part of the intergenic spacer could be bypassed during reverse transcription or PCR, due to RNA or DNA folding. The possibility of such an artefact was confirmed by the large deletion found in a clone (clone 32) which lies between two identical 16 nucleotide sequences (underlined in Fig. 1), one of which is included in the deletion. The other size variation, found in another clone, is a small duplication of nine nucleotides (UUUUUUUAG), inserted between positions 984 and 985 of the consensus sequence. Similar small insertions were found for the intergenic regions of RStV RNAs 3 and 4 (Kakutani et al., 1991; Zhu et al., 1991, 1992).

Sequence variations found once (plain letters) or more than once (bold letters) amongst the clones are also indicated in Fig. 1. There are 11 nucleotide variants scattered through the viral-sense ORF, five of which are silent, two giving conservative amino acid changes and the rest giving non-conservative changes. In the complementary-sense ORF there are 24 nucleotide variants, 13 being silent, one giving a conservative change and the rest giving non-conservative changes; most of these nucleotide variants are in the 3'-terminal part of this ORF. The intergenic region has 15 variants, which are not clustered.

The predicted amino acid sequences of viral-sense and the complementary-sense ORFs of RHBV RNA 3 were aligned with their counterparts in RStV and MStV (Fig. 2). This indicates that the viral-sense ORF product is NS3 and that of the complementary-sense ORF is the N protein; the size of product of the complementary-sense ORF, 35K, is similar to that of the N protein of RHBV (Morales & Niessen, 1983). For the N protein alignment (Fig. 2a) there are two single amino acid insertions and one four amino acid insertion in the RHBV sequence relative to the MStV and RStV sequences. An interesting feature of this alignment is the uneven distribution of the variation between the N proteins of the three viruses. Most of the variation occurs in the C-terminal portion of the protein, which is close to the intergenic spacer of the corresponding RNA. The alignment for the NS3 proteins (Fig. 2b) is much more regular, without insertions and with an even distribution of the variation between the viruses. For both alignments it is apparent that RStV and MStV are much more closely related to each other than they are to RHBV. This is expressed numerically in Fig. 2(c). Both the N and NS3 proteins are about 65% (60% at nucleotide level) identical between RStV and MStV, and about 46% (43% at nucleotide level) identical between RHBV and either RStV or MStV. These figures imply that both proteins have changed at about the same rate, which is unusual given the inevitable functional and structural differences that exist between the proteins.

As expected, the terminal sequences of RHBV RNA 3 mimic those of RStV and MStV RNA 3. The 17 terminal nucleotides at the 5' end of the genomic RNA match...
perfectly with 16 nucleotides at the 3' end. The only difference is the characteristic insertion of a C residue at position 11 for which there is no match at the 3' end (Takahashi et al., 1990; Kakutani et al., 1991). These 17/16 nucleotides are perfectly conserved among the RNAs 3 of RStV, MStV and RHBV, and differ by only two nucleotides (an A at position 12/11 and a C at position 16/15) from the terminal sequences of the RNAs 4 of RStV, MStV and RHBV, which among themselves are also perfectly conserved.

The intergenic regions of the RNAs 3 of the three viruses differ in length (742, 650 and 517 nucleotides for RStV, MStV and RHBV respectively) and in sequence, with only about 30% overall similarity between any two of them. However, closer examination reveals a sequence of about 150 nucleotides at the 5' end of this region (nucleotides 835 to 985 of the RHBV sequence) where there is 55 to 60% identity between the viral sequences, with especially good alignment between RHBV and RStV. A characteristic feature of the intergenic region of the three tenuiviruses are groups of oligo(A) and oligo(U) residues. These are highlighted for RHBV in Fig. 1 where it can be seen that there are more oligo(U) groups than oligo(A) groups. In this, RHBV resembles MStV but differs from RStV, which has more oligo(A) than oligo(U) groups (Kakutani et al., 1991). One further difference is that in RStV most of the oligo(A) groups are in the 5' half of the intergenic region and the oligo(U) groups are in the 3' half. In RHBV and MStV they are much more interspersed.

Apart from the oligo(U) and oligo(A) groups there is a direct repeat of 16 nucleotides in the intergenic region of RHBV RNA 3 (Fig. 1, underlined); with one mismatch each also has an internal repeat. Repeats of 12 to 15 nucleotides were found in the intergenic regions of MStV and RStV but these were made up of adjacent oligo(U) groups.

Most of the viruses that use the ambisense strategy have intergenic regions rich in A and U residues. It has been suggested that they can form hairpin structures that function in transcriptional termination of the ambisense genes (Emery & Bishop, 1987; Schmaljohn & Patterson,
variability is the frequency of variants to the consensus instance, the C/U, A/G and U/C variations at varying lengths of coding RNA, or that the coding RNA This could mean that the predicted folding of the corresponding segments of RStV and MStV RNAs 3. deletion of clone 32 excepted, were unstable with respect to large variations in the amount of coding RNA included. The folding ability of the coding RNA was indistinguishable from that of the intergenic region. Inclusion of coding sequences showed that this also applies to similar computer-assisted folding analyses of the corresponding segments of RStV and MStV RNAs 3. This could mean that the predicted folding of the intergenic region alone is poor enough to be upset by varying lengths of coding RNA, or that the coding RNA is an integral part of the overall folding pattern. If tenuivirus coding RNA does indeed fold extensively, then this would provide a different type of evolutionary constraint, based on RNA secondary structure rather than amino acid conservation, that could help to explain the similar rates of change of the two proteins encoded by this RNA.

The sequence of RHBV RNA 3 is most variable in the intergenic region and least variable at the termini, with the coding regions being of intermediate variability. Although some of the variations found among the clones will have arisen during PCR, the majority are naturally occurring. At about one in four of the variant positions the same difference occurs in more than one clone and for any one clone the variants tended to be clustered. For instance, the C/U, A/G and U/C variations at nucleotides 1232, 1234 and 1236 (Fig. 1) were always found in the same clone. Neither of these phenomena is likely to be due to PCR and the inference is that the viral RNA as a whole is quite variable, which corresponds with data obtained for RStV (Kakutani et al., 1991; Zhu et al., 1991, 1992). However, certain variations are more likely to have been caused by PCR, for example the deletions at nucleotides 653 and 1912 (Fig. 1), which would change the reading frames of the NS3 and N proteins respectively. An approximate estimate of the variability is the frequency of variants to the consensus sequence found among the clones. For the complete sequence this frequency is \(7.4 \times 10^{-5}\) per nucleotide. The least variable elements of the sequence are the termini (4.2 \(\times\) \(10^{-3}\) per nucleotide) followed by the NS3 protein region (5.5 \(\times\) \(10^{-3}\) per nucleotide), the N protein region (7.9 \(\times\) \(10^{-3}\) per nucleotide) and the intergenic spacer (9.7 \(\times\) \(10^{-3}\) per nucleotide). These are all underestimates of the true variability for these regions, since they are all estimated with reference to the consensus sequence which, by definition, is the least variable sequence that can be constructed from the component sequences. However, PCR-introduced errors will have increased these figures slightly, by about 0.4 \(\times\) \(10^{-3}\) per nucleotide (Eckert & Kunkel, 1991) if we assume a Taq polymerase error rate of 2.5 \(\times\) \(10^{-5}\) per nucleotide which is about appropriate for these conditions (Gelfand & White, 1990; Kwiatowski et al., 1991).

Fig. 2(c) shows that RHBV RNA 3 is more closely related to MSIV than to RStV and this also is the case for the RNA 4 genes. Therefore, RHBV cannot be considered to be a South and Central American strain of RStV (which is found in Asia). When comparing the RNA 3 sequences of the three tenuiviruses, both of the encoded proteins differ by about the same amount, with RStV and MSIV about 65% identical, and RHBV about 45% related to either RStV or MSIV. A similar analysis of the amino acid alignments of the two proteins encoded by RNA 4 reveals the same pattern. Both proteins are about 75% identical between RStV and MSIV, and 60 to 60% identical between RHBV and either RStV or MSIV. If this is a general feature of tenuivirus RNAs, then this implies that these RNAs have evolved as an integral unit, without the type of recombinational exchange of genetic modules and evolutionary independence of viral genes that characterizes positive- and negative-strand RNA virus evolution (Strauss & Strauss, 1988; Koonin, 1991; Tordo et al., 1992). Why the proteins should change at a similar rate, as well as why the phylogenetic relationships of the viral proteins should be so symmetrical, with RHBV almost exactly equidistant from both MSIV and RStV, is less easily explained. One possibility is that the RNA of the coding regions, as well as coding for the proteins, has another, possibly more important function, based on the actual RNA sequence and therefore not affected by the triplet code. If this function requires the sequence to adapt rapidly when necessary, in response to positive selection pressure, then the compromise between this selection pressure for change and the conservative selection pressure to maintain protein function could cause the proteins to appear to evolve at the same rate. One obvious candidate for this alternative, RNA-based function is RNA structure (secondary or tertiary), which can be both extensive enough and flexible enough to fulfil such a role and which would control virus expression and/or interactions of the RNA with proteins. However, the evidence that such structures exist to any great extent in the coding RNA of tenuiviruses, or what their function could be, is meagre. In the absence of physical data concerning the folding pattern and potential of tenuivirus RNAs, this must remain conjecture.
Many thanks are due to Dr Gabriel Macaya for help with the analyses and to Dr A.-L. Haenni for providing a preprint of Ramirez et al. (1993). This work was financed by a Commission of European Communities Joint Research grant and a Rockefeller Foundation grant.

References


(Received 15 September 1993; Accepted 3 March 1994)