An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus

Joanne M.Hay*, Matthew C.Jones+, Maggi L.Blakebrough, Indranil Dasgupta, Jeffrey W.Davies and Roger Hull
Department of Virus Research, John Innes Institute, John Innes Centre for Plant Science Research, Colney Lane, Norwich NR4 7UH, UK

Received March 4, 1991; Revised and Accepted April 23, 1991 EMBL accession no. X57924

ABSTRACT

The nucleotide sequence of an infectious clone of rice tungro bacilliform virus (RTBV) DNA has been determined. The circular genome has 8002 bp and one strand contains four open reading frames (ORFs). One ORF is potentially capable of encoding a protein of 24 kD (P24) and has no initiation (ATG) codon. The other three ORFs potentially encode proteins of 12 kD, 194 kD and 46 kD (P12, P194, P46) respectively. The functions of P24, P12 and P46 are unknown. Comparative analyses with retroviruses and *Commelina* yellow mottle virus suggest that the 194 kD putative product is a polyprotein that is proteolytically cleaved to yield the virion coat protein, a protease and replicase (reverse transcriptase and RNase H) characteristic of retroelements. The DNA sequence reveals other features which strongly support our belief that RTBV is a pararetrovirus. These include sequences at the mapped positions of two discontinuities in the virion DNA which are complementary to tRNA $^{met}{}_{init}$ and purine-rich, and may be the priming sites for minus- and plus-strand DNA synthesis respectively. As the positions of likely transcriptional signals suggest, a full-length viral transcript is observed by northern analysis. The predicted folding of the 645 bp 5′-region of this RNA resembles that of caulimoviruses. Comparisons with other reverse transcribing elements are discussed.

INTRODUCTION

Rice tungro disease is the most economically important viral disease of rice and causes annual losses in South and South-East Asia estimated at $\$1.5\times10^9$ (1). A complex comprising rice tungro spherical virus (RTSV) and rice tungro bacilliform virus (RTBV), is responsible for the disease (2, 3). RTSV infection causes mild or indistinct symptoms and is leafhopper-transmitted (principally by *Nephotettix virescens)* (3). The particles are isometric, have a diameter of 30nm and contain a single-stranded RNA genome of more than 10 kbp (4). RTBV infection in rice causes yellowing of leaves and results in stunted growth; the symptoms are accentuated in plants co-infected with RTSV (5). The bacilliform particles have a modal length of 130nm and width of 30nm (3) and can only be acquired from plants by leafhoppers in association with RTSV (5).

RTBV is a proposed member of the recently recognised group of viruses with non-enveloped bacilliform particles (6), termed badnaviruses (*ba*cilliform *DNA viruses*). RTBV and group members infecting *Kalanchoe*, cocoa, *Commelina*, and sugarcane have been shown to contain circular double-stranded DNA genomes (4, 7-9). The RTBV genome is 8.0 kbp and is interrupted by two discontinuities which map at specific sites, one in each strand (4). Neither RTBV nor its DNA is transmissible mechanically into rice plants. Infectivity of an RTBV clone has been shown by agroinfecting constructs of more than viral genome length into rice (10).

The only other group of plant viruses with circular double-stranded DNA genomes is the caulimoviruses (11), which are pararetroviruses (12). Like RTBV, caulimovirus genomes are approximately 7.8–8.0 kbp and have site-specific discontinuities, one in the minus strand and one or more in the other (13). However, caulimoviruses have isometric particles, their genomes encode at least six proteins (14) and do not cross-hybridise with badnavirus genomes (4).

The replication of cauliflower mosaic virus (CaMV) is biphasic and the viral DNA genome is transcribed in the nucleus to give a greater-than-genome length RNA which is the template for reverse transcription in the cytoplasm (for review see 14). A replicase (including protease, reverse transcriptase and RNase H domains) encoded by caulimovirus gene V is thought to be involved in the cytoplasmic phase and mapping at the site-specific discontinuities suggest that the cytosolic initiator methionine tRNA ($tRNA^{met}{}_{init}$) and purine-rich oligonucleotides prime reverse transcriptase directed minus- and plus-strand synthesis, respectively.

In a recently published sequence, the DNA of the type member of the badnavirus group, *Commelina* yellow mottle virus (CoYMV) was shown to have a genome of 7489 bp which

* To whom correspondence should be addressed

+ Present address: Department of Biochemistry, University of Cambridge, Cambridge, UK potentially encoded three polypeptides (7). The largest polypeptide, with a Mr of 216,000 (P216), had motifs characteristic of the coat protein (*gag*), and of polymerase (*pol*). The nucleic acid sequence at two discontinuities indicated that they could be the sites for priming DNA synthesis. Medberry *et al.* (7) also reported that CoYMV DNA was transcribed to give a more-than-genomic length RNA which had a direct terminal repeat of 120 nucleotides.

We have determined the sequence of an infectious clone of RTBV DNA. Analysis of the sequence has identified four open reading frames (ORFs) capable of encoding proteins of more than 100 amino acids. One of the ORFs has no potential initiating ATG codon. Comparative analysis suggests that the largest ORF encodes a polyprotein which may be processed to yield the viral coat protein, and the protease, reverse transcriptase and RNase H characteristic of retroelements. The sequence in the region of one of the discontinuities is homologous to the cytoplasmic $\mathrm{tRNA^{met}}_{\mathrm{init}}$ which suggests that, by comparison with caulimoviruses, this region primes reverse transcriptase-directed minus-strand DNA synthesis.

MATERIALS AND METHODS

Virus isolate and clones

Tungro (RTBV and RTSV) (Philippine isolate, kindly donated by Dr. H. Hibino, International Rice Research Institute) was propagated in rice (*Oryza sativa*) var. TN1 and the virus particles and virion DNA purified as described (4). Genomic clones were constructed by ligating RTBV DNA digested with either *SalI* or *BamHI* into pUC18. Two full-length clones, pJIIS2 (*SalI*) and pJIIB4 (*BamHI*), were identified by restriction mapping (4) and pJIIS2 was shown to be infectious (10).

DNA sequencing

Single-stranded sub-fragments of viral DNA from pJIIS2 and pJIIB4, as well as the complete double-stranded plasmids, were used as sequencing templates. Restriction fragments were obtained from the purified plasmids and inserted into the appropriate linearized and alkaline-phosphatased restriction site of the M13 vectors mp18 and mp19 (15). Recombinant phages were identified by the lac complementation assay (16); bacteriophage isolation and DNA extraction were as described by Sambrook *et al.* (16). Plasmid DNA was denatured in preparation for double-stranded sequencing in 0.2M sodium hydroxide and 0.2M EDTA at 37°C for fifteen minutes and precipitated in 100% ethanol and 2M ammonium acetate, pH 4.5.

Sequencing was performed using [α-^{35}S]dATP and a Sequenase kit (US Biochemical) as recommended by the manufacturer. The products were electrophoresed on a 6% (w/v) polyacrylamide gel (17) and were fixed, dried and subjected to autoradiography. Either a 17-nucleotide universal sequencing primer (US Biochemical) or 15–18-nucleotide primers complementary to appropriate regions of the viral DNA were used to prime the sequencing reactions for both single-stranded and double-stranded sequencing. Nucleic acid sequence data were assembled and analysed using programs from the University of Wisconsin Genetics Computer Group (UWGCG; 18) and the Staden package (19).

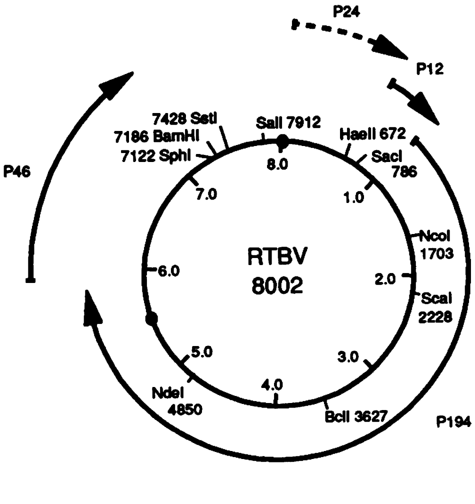


Figure 1. Potential protein coding regions (arrows) designated P24, P12, P194 and P46 are identified within RTBV DNA in the virion DNA plus strand. The dashed arrow of P24 shows that no ATG codon was identified. The inner circle represents the viral DNA. . indicates sites of the discontinuities and the position of some unique restriction sites are indicated.

RNA isolation and northern analysis

Total RNA was extracted from infected and healthy leaves in glasshouse grown *O. sativa* var. TN1 using a modification of the method of Leaver and Ingle (20). Plant material was powdered in liquid nitrogen and thawed in one volume (v/w) of extraction buffer (1% tri-isopropyl-naphthalene sulphonate, 6% para-amino salicylate, 50mM NaCl, 10mM Tris pH7.4). The ground material was extracted twice with an equal volume of 1:1 (v/v) phenol:chloroform. The aqueous phase was mixed with two volumes of ethanol and the nucleic acids collected by centrifugation. The pellet containing RNA was resuspended in 0.1M sodium acetate buffer (pH5.0), containing 0.5% sodium lauryl sulphate and 1% butanol. RNA was reprecipitated with two volumes of ethanol.

Total polyadenylated [poly(A)] RNA was enriched using an oligo(dT)-cellulose protocol (16). RNA was denatured, electrophoresed under denaturing conditions using formaldehyde and formamide and transferred to Hybond N (Amersham) by standard techniques (16). A [^{32}P-dCTP] labelled RTBV-specific hybridisation probe (pJIIS2) was prepared by random-priming (21), and used for northern analysis as recommended by Sambrook *et al.* (16).

RESULTS

The sequence

The DNA sequence of pJIIS2 (EMBL accession number X57924) shows that the infectious clone of RTBV has a circular genome of 8002 bp. The complete genome was covered by sequence from

Table 1. Protein coding regions of RTBV genome

ORF designation	Reading frame	Starting nucleotide*	First ATG**	Ending nucleotide***	Molecular weight
P24	+3	69	–	668	23,692^{+}
P12	+2	653	665	997	11,911^{++}
P194	+1	979	994	6021	194,083^{++}
P46	+3	5945	6042	7211	46,173^{++}

* Starting nucleotide of ORF; ** A of ATG start codon; *** 3rd base of stop codon;

$^{+}$ MW calculated from start of ORF and not ATG; ++ MW calculated from first ATG both strands and much of it was confirmed by sequencing pJIIB4. By analogy with published caulimovirus and CoYMV sequences, the numbering of the RTBV sequence begins at the 5′ end of the putative minus-strand priming site (see below). The RTBV genome has a G + C content of 33.7% which is lower than that of CoYMV (39.6%; 8) and of most caulimoviruses whose G + C contents range from 34.0% for soybean chlorotic mottle virus (SCMV) (22) to 44% for figwort mosaic virus (FMV) (23). A comparison of the sequences derived from pJIIB4 and pJIIS2 identified four nucleotide variations: (74 A or G, 81 G or A, 5875 G or A, 7981 A or G). The substitutions of an A residue at position 81 and a G residue at position 5875 would result in amino acid changes of D to N and E to K, respectively. Neither of the other two changes would result in any amino acid change or introduction of stop codons. It is considered that these differences reflect the heterogeneity of the RTBV virion DNA and do not represent cloning artifacts.

Analysis of the sequence showed that 29 restriction endonucleases digested the RTBV DNA at one site only. The map positions of some which might be useful in manipulating RTBV DNA are shown in Fig 1. The restriction endonuclease sites predicted from the sequence showed agreement with the map positions of those reported by Jones *et al.* (4), except for the number and position of some *Xba*I sites, an additional *Acc*I site at 6549 bp and a further *Cla*I site at 7406 bp. Attempts to map this latter site in viral and cloned DNA failed which suggests that it is easily methylated.

Coding regions

The coding capacity of the RTBV DNA was investigated and the sequence screened in all reading frames for potential genes using the program FRAMES (18). Four ORFs capable of encoding proteins larger than 10 kD were identified in the strand containing the minus-strand priming site (see below). The largest ORF in the complementary strand (from 4428–4700 in that strand) is capable of encoding a polypeptide of 91 amino acids, but the first methionine is the 47th amino acid. The ORFs in the plus strand constitute 90% of the genome; their relative positions are presented on a circular map in Fig 1 and details of them are given in Table 1.

The first ORF of the sequence is a long region with no potential ATG initiation codon. It overlaps the second ORF by a frame change of −1 with the sequence ATGA, the TGA being the stop codon of the first ORF and the ATG being the first initiation codon of the second ORF. This resembles sequence motifs found in some caulimoviruses especially carnation etched ring virus (CERV) (24). The same sequence motif is found at the interface of RTBV second and third ORFs but the stop codon of the third ORF (TAA) is separated from the first initiation codon of the fourth ORF by 20 nucleotides.

Comparison of amino acid sequences

Computer analyses were performed to detect any similarities between the putative products of the four RTBV ORFs (P24, P12, P194 and P46) and the proteins contained in the EMBL and NBRF data bases using the program TFASTA (25). No significant homologies were found for P24, P12 or P46. Using the program SCRUTINEER (26), which searches a data base of motifs, a homology between amino acids 427 to 443 of P194 with the active site of the Kunitz-type trypsin inhibitor family (27) was detected (Fig 2A). The significance of this homology is not known. The DIAGON program (19) (span length 31, score

```
A

Consensus        hXDXXGXXhXXXXXYXh
RTBV P194 427   -IEDIFGELLKEHGNYDM-
                 : *  *  :     * :

B

RNA binding domain                                  Protease

RTBV  P194 772   CRCYICQDENHLANRCPR                 RTBV  P194  986  IDSGSTHNIICPTLIPA
CoYMV P216 879   CKCYICGQEGHYANQCRN                 CoYMV P216 1219  VDTGATACLIQISAIPE
                 *:**** :* * ** *                                    :*:*:*  :*  : **
CaMV ORFIV 409   CRCWICNIEGHYANECPN                 CaMV ORFV    44  VDTGASLCIASKFVIPE
CERV ORFIV 417   CRCWVCNIEGHYANECPN                 CERV ORFV    32  VDTGSSLCMASKYVIPE
FMV  ORFIV 408   CRCWICTEEGHYANECPN                 FMV  ORFV    52  VDTGASLCIASRYIIPE
SCMV ORFIV 380   CQCWLCHEEGH-ANTCPK                 SCMV ORFV    35  IDTGATLCFGKRKISNN
                 * *::*  *"* ** *"                                   :*:*:: "      ""

Reverse transcriptase

RTBV  P194 1274   IFSKFDLKAGFHHM <20> WNVCPFGIANAPCAF <11> KFALLYIDDILIAS
CoYMV P216 1498   IYSKFDLKSGFWQV <20> WLVMPFGLKNAPAIF <12> KFIAVYIDDILVFS
                  *:******:**  :      * * ***: ***  *      **  :******: *
CaMV ORFV   335   IFSSFDCKSGFWQV <20> WNVVPFGLKQAPSIF <12> KFCCVYVDDILVFS
CERV ORFV   316   IYSSFDCKSGLWQV <20> WNVVPFGLKQAPSIF <13> KYCCVYVDDILVFS
FMV  ORFV   327   IFSSFDCKSGFWQV <20> WKVVPFGLKQAPSIF <12> KFCMVYVDDIIVFS
SCMV ORFV   291   WFSSLDAKSGYYQL <21> WNVLSFGLKQAPCIY <11> DHILAYIDDILIFT
                  ":*""* *:*  ":      * * "**:":** ":      "    *:***::":

Ribonuclease H

RTBV  P194 1486   IIETDASEEGWG <15> EKIAGYASGNFG <72> EHIKGNKNFLPNFL
CoYMV P216 1711   IIETDGCMTGWG <15> ERICAYASGSFN <72> EHIDGKHNGLADAL
                  *****:   ***      *:* :**** *       *** * :* *:: *
CaMV ORFV   547   IIETDASDDYWG <14> ELICRYASGSFK <72> EHIKGTDNHFADFL
CERV ORFV   530   VIETDASEEFWG <10> EYICRYASGSFK <72> EHIAGTKNVFADFL
FMV  ORFV   539   IIETDASDSFWG <11> ELICRYSSGSFK <72> EHLEGVKNVLADCL
SCMV ORFV   508   IVETDASQHSWS <63> LLLCKYVSGTFT <72> ELIKSENNPFEIRL
                  ::***:"   *"      " :" * ** *       *": :  * "   *
```

Figure 2. Comparison between RTBV P194 and (A) Legume Kunitz-type trypsin inhibitor consensus amino acid sequence (26); h = small hydrophobic amino acid, L,I,V,M; X= any amino acid; (B) domains from CoYMV P216 (7) and gene V products of CaMV (11), CERV (24), FMV (23) and SCMV (22) (for virus abbreviations see text). * indicates direct homology, : familial homology and ″ two unrelated groups of direct homology.

345) was used to compare the predicted products of RTBV ORFs with those of CoYMV (not in data bases). A short homology was found between RTBV P24 and CoYMV P15 (7) but its significance is not known. No homologies were found between RTBV P12 or P46 and the proteins suggested to be encoded by CoYMV. Analysis of RTBV P194 using TFASTA revealed several sequence motifs characteristic of retroelements. These included the 'cys' RNA binding domain of caulimovirus coat proteins and the nucleocapsid portion of *gag* proteins (28), and the aspartic protease, reverse transcriptase and RNase H domains of retroviruses, caulimoviruses and other reverse transcribing elements (Fig 2B) (29).

Medberry *et al.* (7) concluded that CoYMV P216 was a polyprotein comprising the coat protein and viral polymerase and the same appears to be so for RTBV P194. Comparison of RTBV P194 with CoYMV P216 by DIAGON (Fig 3A) and GAP (18) showed several regions of homology. Region 1 (26% direct homology, 46% direct + familial homology − designated 26%/46%) is approximately between amino acid positions 200 and 400 on RTBV P194 and showed no motifs or homologies with recorded functional proteins. It is, however, somewhat basic with 30 basic amino acids and 20 acidic ones. Region 2, (22%/42% homology) between amino acids 570 and 820, contained the 'cys' sequence $CX_2CX_4HX_4C$ and the rest of the region was relatively basic. Region 3 (48%/70% homology), between amino acids 1180 and 1610, contained homologies with reverse transcriptase and ribonuclease H domains. Homology in the potential protease domains was not detected at the stringency used for this DIAGON. The spatial arrangement of these homologies on RTBV P194 and CoYMV P216 is shown in Fig 3B. This can be compared with Fig 6 of Medberry *et al.* (7) which relates the spatial arrangement of these domains in CoYMV and CaMV.

DIAGON analysis showed that the strongest homology between caulimovirus coat proteins and RTBV P194 is in region 2, the 'cys' sequence plus adjacent basic sequence described above. There is no evidence of N- and C-terminal regions rich in acidic amino acids as are found in CaMV coat protein (11). DIAGON analysis of the polymerase domain of RTBV P194 showed that, as with the comparison with CoYMV P216, the homologies with caulimovirus gene V products were in RTBV region 3. The region of P194 encoding coat protein motifs showed less

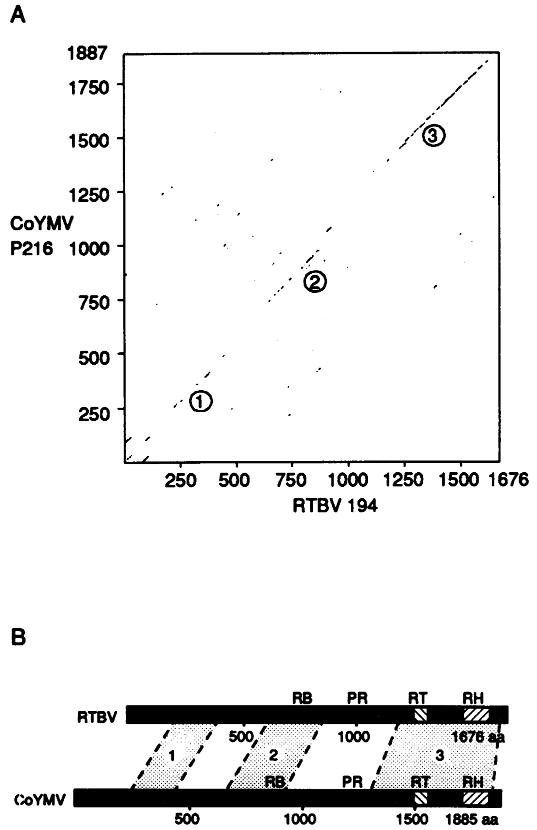


Figure 3. (A) Comparison of RTBV P194 and CoYMV 216 by DIAGON. Three major regions of homology were identified as indicated. (B) Comparison of the spatial distribution of various functional domains in RTBV P194 and CoYMV P216 (7). RB = RNA binding domain, PR = protease domain, RT = reverse transcriptase domain, RH = RNase H domain, aa = amino acids. Also shown are the regions of homology identified in (A).

Table 2. Sequence homologies between RTBV, CoYMV and caulimoviruses

Minus-strand primers

													*					*		
tRNA $^{met}_{init}$ consensus[1]		A	C	C	A	U	A	G	U	C	U	C	G	G	U	C	C	A	A	
RTBV	1	T	G	G	T	A	T	C	A	G	A	G	C	g[2]	A	t	G	T	T	
CoYMV	1	T	G	G	T	A	T	C	A	G	A	G	C	t	t	G	G	T	T	T
Caulimoviruses																				
CaMV	1	T	G	G	T	A	T	C	A	G	A	G	C	C	A	t	G	a	a	
CERV	1	T	G	G	T	A	T	C	A	G	A	G	C	C	A	t	a	g	T	
FMV	1	T	G	G	T	A	T	C	A	a	A	G	C	C	A	t	G	T	g	
SCMV	1	T	G	G	T	A	T	C	A	G	A	G	C	a	A	G	a	T	T	

Plus-strand primers

RTBV	5679	AATAGTGAAGGGGATAAAA
CoYMV	4693	GGGGTAAGAAGGGGAAGG
CaMVG2	4208	AGAGGGGAGGAGG
CaMVG3	1622	AAGAGTGGGGGGG
CERVG2	3868	AGAAGAGGGGGGGAA
CERVG3	1265	AAGGAGGGGGGAGGAA

TATA and polyadenylation signals

	TATA		Poly(A) signal	
RTBV	7372	GTATATAA	7598	AATAAA
CoYMV	7321	CTTATTTAA	7454	AATAAAA
CaMV35S	7403	CTATATAA	7598	AATAAAA
CERV35S	7050	CTATATAA	7247	ATAATTA
FMV35S	6895	CTATATAA	7064	AATAAA

[1] tRNA $^{met}_{init}$ consensus sequence is of plant cytoplasmic tRNAs (30−32); [2] lower case letters refer to non-complementary amino acids; * methylated amino acids; sequence references are: CoYMV (7); SCMV (22); CERV35S, gap2 (CERVG2), gap3 (CERVG3) and CERV minus-strand primer sequences (24); CaMV 35S, gap2 (CAMVG2), gap3 (CAMVG3) and CAMV minus-strand sequences (11); FMV35S and FMV minus-strand primers (23).

homology with caulimovirus coat proteins, than the P194 region (943–1676 amino acids) encoding the putative protease and polymerase had with caulimovirus polymerases.

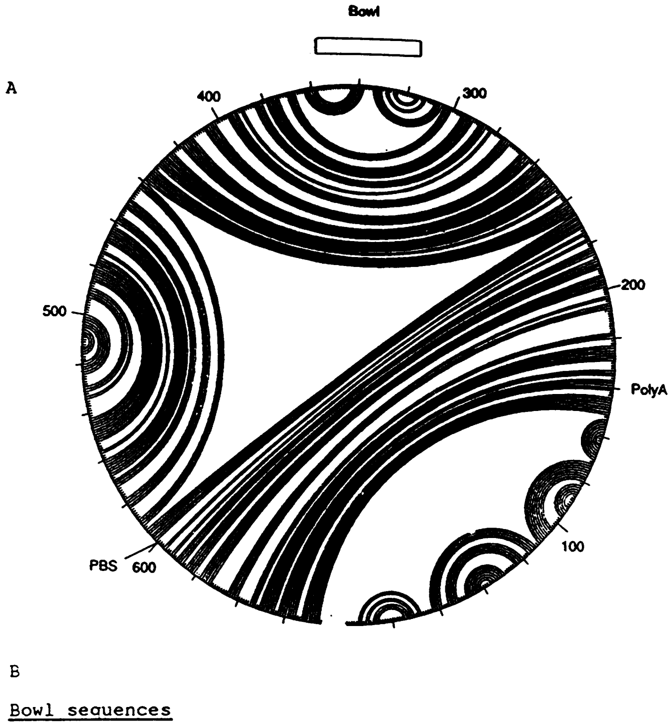
Putative primer binding sites

Analysis of the RTBV sequence around the positions of the two gaps mapped by Jones *et al.* (4) revealed a sequence on the plus-strand DNA, with complementarity to the first fourteen 3′-terminal nucleotides of the $tRNA^{met}{}_{init}$ from plant cytoplasms (Table 2) (30–32). The cytosolic $tRNA\ ^{met}{}_{init}$ is also considered to be the primer for the synthesis of minus-strand DNA of CoYMV (7) and caulimoviruses (22–24).

The priming sites for the synthesis of the plus-strand DNA of retroelements are usually polypurine regions (33). There is a purine-rich sequence near the site where the second discontinuity of RTBV DNA has been mapped (Table 2) (4). This shows some homology to the sequence identified as the plus-strand priming site of CoYMV (7).

Non-coding regions

There is a large intergenic region between the ORFs encoding P46 and P24 (nucleotides 7212–69) (Fig 1). No sequence homology, other than the $tRNA^{met}{}_{init}$ complementary sequence, was found when the large intergenic regions of RTBV and CoYMV were compared by DIAGON. As noted earlier, the first initiation codon of the P46 ORF is separated from the P194 ORF stop codon by 20 nucleotides. However, the P46 ORF overlaps the P194 ORF by 67 nucleotides and as it could be expressed by mechanism such as frame-shift or relay race (34) these ORFs may not be separated by an intergenic region.



B

Bowl sequences

```
CaMV 7727  AAGCTAGAAGTACCGCTTAGGCAGGAGGCCGT-TAGGGAAAAGATGC
FMV  7220  GGATCTGAAGTACCGCCGAGGCAGGAGGCCGT-TAGGGAAAAAGGGA
CERV 7339  AAGGGTG-AGTACCGCCGAGGCAGGAGGCCGTATAGGGAAAACAGGT
                 : ::::::::  :::::::::::::: ::::::::   :
RTBV 7708  GGGGAAAAAGTACCG-TCAGGCCGTGTTATGGCAAGGGAAGAAGTAC
                   ++++++++  ++++ +      +  ++++++ +
```

Figure 4. Secondary structure features of the first 645 nucleotides of the proposed RNA transcript of RTBV. (A) CIRCLE presentation derived from the FOLD program with the sites of the suggested polyadenylation consensus sequence (Poly A), the 'bowl' sequence (see text) and the minus-strand DNA primer site (PBS) indicated. (B) Alignment of the 'bowl' sequences of RTBV and those of CaMV, FMV and CERV. : indicates common nucleotides in all three caulimoviruses and + indicates nucleotides common to all four sequences.

In the RTBV sequence a putative TATA box sequence (Table 2) is found in the large intergenic region at position 7372 and closely resembles those of caulimovirus 35S RNA, but differs from that of CoYMV (7). The polyadenylation signal in caulimoviruses is found about 200 nucleotides downstream of the TATA box. In RTBV a sequence resembling the consensus polyadenylation signal (35) is found at nucleotide 7598 (Table 2) and is similar to signals reported for CoYMV and caulimoviruses. The octopine synthase (ocs) enhancer element sequence, found in the promoter regions of some caulimoviruses (36), was not detected in the RTBV large intergenic region.

The leader sequence of the caulimovirus large (35S) RNA transcript can be folded into a characteristic large stem-loop structure (37). The folding of the first 645 nucleotides (7396–39) (Fig 4A) from the putative promoter of the RTBV large RNA transcript strongly resembles that of the caulimoviruses (compare Fig 4A with Fig 1 of ref. 37). Fütterer *et al.* (37) noted that the feature of the folded caulimovirus leader sequence they termed the 'bowl' was the conserved sequence within caulimoviruses noted by Richins *et al.* (23). The 'bowl' sequence of RTBV (nucleotides 7708–7753) has a close resemblance to those of caulimoviruses (Fig 4B), but this sequence was not found in CoYMV.

Northern analysis

Northern blots of total RNA from RTBV-infected rice plants showed an RNA species of about 8 kb (Fig 5, lane B). Two smaller RNA species of 3.1 kb and 0.5 kb hybridised with the RTBV probe at a much lower intensity. The smear of

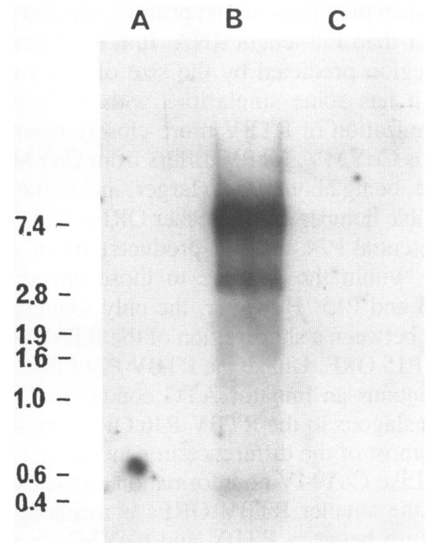


Figure 5. Northern blots of total RNA from RTBV infected and healthy *O. sativa* var. TN1. All lanes were hybridised with a ^{32}P-labelled pJIIS2 probe. Lane (A) was loaded with 25 μg of total RNA from healthy plants; (B) was loaded with 25 μg total RNA from infected plants and (C) was loaded with 10 μg of polyadenylated RNA from infected plants. MW ($\times 10^3$) are indicated.

hybridisation below the 8.0 kb transcript is probably due to degradation of the larger transcript and could obscure detection of less abundant transcripts. No transcripts from RNA extracted from healthy plants hybridised with the virus-specific probe (Fig 5, lane A), indicating the transcripts in lane (B) were virus-encoded. The three virus-encoded transcripts appear not to be polyadenylated (Fig 5, lane C).

DISCUSSION

Sequencing of an infectious clone has shown that the genome of RTBV has 8002 bp, slightly smaller than the size predicted by restriction endonuclease mapping (4). The sequence data, together with the properties of the RTBV genome reported by Jones *et al.* (4), strongly suggest that RTBV is a pararetrovirus and that its replication includes a reverse transcription phase. It resembles other plant pararetroviruses in having double-stranded genomic DNA containing site-specific discontinuities which are likely to be the sites for priming DNA synthesis. It also potentially encodes a polymerase which may be capable of reverse transcription, but lacks an endonuclease domain. The putative minus-strand DNA priming site is complementary to the 3′-terminus of $tRNA^{met}{}_{init}$ which is the suggested priming site for all plant retroviruses and retroelements. The major transcript appears to be non-polyadenylated and full length, and therefore resembles the major transcript from CoYMV (7). As with caulimoviruses, this transcript could be the template for reverse transcription. In contrast with CoYMV neither of the smaller virus-specific transcripts are found in healthy samples (Fig 5); these transcripts are currently being investigated. Observations supporting a reverse transcription mode of replication were agarose gel electrophoresis and Southern blotting of total DNA from RTBV-infected rice plants which revealed virus-specific DNA species of about 500–600 nucleotides (unpublished observation). This is of a size similar to the small 'strong stop' DNA of CaMV (38) formed between the minus-strand priming site and the 5′ end of the 35S RNA, and which delimits the distance between the minus-strand priming site and the promoter for the greater-than-full-length RNA. In RTBV there is a TATA box in the region predicted by the size of this small DNA.

Although it has some similarities with caulimoviruses, the genome organization of RTBV more closely resembles that of the badnavirus CoYMV. RTBV differs from CoYMV in the size of its genome, being about 0.5 kb larger, and in the distribution, size and possible number of the smaller ORFs. Two RTBV ORFs (encoding potential P24 and P12 products) are of a similar size and position within the genome to those encoding CoYMV products P23 and P15. However, the only sequence homology detected was between a short region of the RTBV P24 ORF and the CoYMV P15 ORF. Unlike the RTBV P24 ORF, the CoYMV P23 ORF contains an initiator ATG codon. In CoYMV there is no ORF analagous to the RTBV P46 ORF, the size of which accounts for most of the difference in genome size between the two viruses. Like CoYMV no information on the expression and functions of the smaller RTBV ORFs is available. The major common feature between RTBV and CoYMV is a large ORF which appears to encode a polyprotein containing the viral coat protein, and the protease, reverse transcriptase and RNase H domains of the polymerase. Medberry *et al.* (7) discussed the significance of this ORF in CoYMV and pointed out that the region potentially encoding the coat protein was much larger than any coat protein species that they observed. The same observations were made for RTBV with the region upstream of the protease domain capable of encoding a protein of more than 1000 amino acids (more than 100 kD)(Fig 3B). The largest coat protein species observed for RTBV is about (70 kD) (Fig 1 in ref 4). We observed that region 1 of the amino acid homology between RTBV P194 and CoYMV P216 is in the N-terminal region and is rather basic. This resembles the product of CaMV gene III, located upstream of the viral coat protein gene (IV), which is also basic and is found in virus particles.

Although the first RTBV ORF (encoding P24) has no initiation ATG codon or any other methionine codons, it would be surprising if a potential coding sequence of 599 nucleotides was not expressed. Furthermore, it overlaps the following ORF (encoding P12) with an ATGA motif which is found at the inteface of CoYMV P23/15 and of several caulimovirus genes. The P24 ORF does not encode the rare initiation codons ACG or GTG noted by Kozak (39), and there is no upstream reading frame which could initiate the expression of P24 by frame shift. It is separated by an ochre stop codon (TAA) from a short in-frame ORF (potentially encoding 23 amino acids) which contains an ATG codon. Although the ochre stop codon is generally regarded as a 'strong' stop, Ishikawa *et al.* (40) provided evidence that it can be read through; a potential 'read-through P24' product would have a MW of 26, 411. A further possibility is the P24 ORF could be spliced to a reading frame further upstream.

The leader sequence for the full-length RNA of RTBV resembles that of caulimoviruses in its characteristic predicted fold structure and also shares sequence homology with a region common to caulimoviruses, termed the 'bowl' sequence (Fig 4B). Fütterer *et al.* (37) suggested that this apparently exposed looped motif might interact with cellular factors in the control of translation. Subsequently, it was reported that caulimovirus gene VI product, which encodes the virus inclusion body protein and is translated from a separate mRNA, was involved in translational transactivation (41, 42) but it is not known if it interacts with the 'bowl' sequence. Inclusion bodies are sites of virus accumulation and are thought to be sites of the reverse transcription phase of virus replication (43). Like CoYMV, no virus-specific inclusion bodies have been observed in RTBV-infected tissue, but in view of the similarities between CaMV 35S and RTBV full-length RNA leader sequences, it would be interesting to determine if any RTBV gene products can also transactivate translation.

Two major groups of reverse transcribing elements have been recognized based on homologies within the *pol* gene products (29). The main group contains retroviruses, the caulimoviruses and the *gypsy* element from *Drosophila*. The other group contains the yeast Ty element and the *Drosophila copia* element. The *pol* gene of retroviruses and retroelements is usually expressed as a polyprotein with the *gag* gene, and in most cases expression involves a frame shift or reading through a weak termination codon (14). CaMV is thought to be different because evidence suggests that the *pol* gene (V) is expressed independently from the *gag* gene (IV) (44). Some plant retroelements express their *gag* and *pol* genes as a single polyprotein (45–48). The *pol* from tobacco, pea and *Arabidopsis* retroelements (45–47) show close affinity to the *copia*-type elements, whereas the *pol* from lily retroelements is more closely related to the *gypsy*-type elements (48). Diagon analysis showed that RTBV P194 had much closer affinities to the *gypsy* than to the *copia pol* sequence (data not shown).

Although reverse transcribing sequences have some common features (eg. *gag* and *pol* genes) there are several areas of divergence between them. These differences include the integration of retroelements into the host DNA, whereas pararetroelements replicate episomally. In most elements the *gag* and *pol* genes are encoded as separate ORFs, but in some retroelements these are encoded on a single polypeptide. There is also a difference in the consensus sequences between the *pol* of the *copia*-type and *gypsy*-type of retroelements. Further more all the plant reverse transcribing sequences use $\mathrm{tRNA}^{\mathrm{met}}_{\mathrm{init}}$ as a minus-strand primer whereas reverse transcribing sequences from animals use a variety of tRNAs (49). An investigation into these divergencies could help elucidate the evolution of this important class of nucleic acids.

ACKOWLEDGEMENTS

We thank Drs C.Poggi-Pollini, J.Moseley and Miss J.Cliffe for assistance with sequencing; Mr M.Harvey for synthesising sequencing oligonucleotides and for preparing figures; Mr A.Lucy for synthesising oligonucleotides and Dr G.P.Lomonossoff for helpful comments on the manuscript. J.M.H was funded by the ODA grant and M.C.J., M.L.B., I.D. were funded by the Rockefeller Foundation. RTBV was held under MAFF licence no. PHF 48A/116 and genetic manipulations carried out under MAFF licence no. PHF 1185A/74(21).

REFERENCES

1. Herdt,R.W. (1988) *Development: Seeds of Change*, **4**, 19–24.
2. Saito,Y., Iwaki,M. and Usugi,T. (1976) *Ann. Phytopath. Soc. Japan*, **43**, 375.
3. Hibino,H., Roechan,M. and Sudarisman,S. (1978) *Phytopath*, **68**, 1412–1416.
4. Jones,M.C., Gough,K., Dasgupta,I., SubbaRao,B.L., Cliffe,J., Shen,P., Kaniewska,M., Blakebrough,M., Davies,J.W., Beachy,R.N. and Hull,R. (1991) *J. Gen. Virol.*, **72**, 757–761.
5. Hibino,H. (1983) *Pl. Dis.* **67**, 774–777.
6. Lockhart,B.E.L. (1990) *Phytopath.*, **80**, 127–131.
7. Medberry,S.L., Lockhart,B.E.L. and Olszewski,N.E. (1990) *Nucleic Acids Res.*, **18**, 5505–5513.
8. Lockhart,B.E.L. and Khaless,N. (1988) *Phytopath.*, **78**, 1548.
9. Lockhart,B.E.L., Bouhida,M. and Olszewski,N.E. (1988) *Phytopath.*, **78**, 1559.
10. Dasgupta,I., Hull,R., Eastop,S., Poggi-Pollini,C., Blakebrough,M., Boulton,M.I. and Davies,J.W. (1991) *J. Gen. Virol.*, (in press).
11. Franck,A., Guilley,H., Jonard,G., Richards,K.E. and Hirth,L. (1980) *Cell*, **21**, 285–294.
12. Temin,H.M. (1989) *Nature*, **339**, 254.
13. Volovitch,M., Drugeon,G. and Yot,P. (1978) *Nucleic Acids Res.*, **5**, 2913–2925.
14. Mason,W.S., Taylor,J.M. and Hull,R. (1987) *Adv. Virus Res.*, **32**, 35–96.
15. Messing,J. and Vieira,J. (1982) *Gene*, **19**, 269–276.
16. Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular cloning: a laboratory manual* (2nd edition). Cold Spring Harbour Laboratory Press, Cold Spring Harbor, NY.
17. Sanger,F. and Coulson,A.R. (1978) *J. Mol. Biol.*, **94**, 441–448.
18. Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.*, **12**, 387–395.
19. Staden,R. (1982) *Nucleic Acids Res.*, **10**, 2951–2961.
20. Leaver,C.J. and Ingle,J. (1971) *Biochem. J.*, **123**, 235–243.
21. Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.
22. Hasegawa,A., Verver,J., Shimada,A., Saito,M., Goldbach,R., van Kammen,A., Miki,K., Kameya-Iwaki,M., and Hibi,T. (1989) *Nucleic Acids Res.*, **17**, 9993–10013.
23. Richins,R.D., Scholthof,H.B. and Shepherd,R.J. (1987) *Nucleic Acids Res.*, **15**, 8451–8466.
24. Hull,R., Sadler,J. and Longstaff,M. (1986) *EMBO J.*, **5**, 3083–3090.
25. Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
26. Sibbald,P.R. and Argos,P. (1990) *CABIOS*, **6**, 279–288.
27. Mares,M., Meloun,B., Pavlik,M., Kostka,V. and Baudys,M. (1989) *FEBS Lett.*, **251**, 94–98.
28. Covey,S.N. (1986) *Nucleic Acids Res.*, **14**, 623–633.
29. Doolittle,R.F., Feng,D.-F., Johnson,M.S. and McClure,M.A. (1989) *Qu. Rev. Biol.*, **64**, 1–30.
30. Canaday,J., Guillemaut,P. and Weil,J. (1980) *Nucleic Acids Res.*, **8**, 999–1008.
31. Ghosh,H.P., Ghosh,K., Simsek,M. and RajBhandary,U.L. (1982) *Nucleic Acids Res.*, **10**, 3241–3247.
32. Sprinzl,M., Hartmann,T., Meissner,F., Moll,J. and Vorderwulbecke,T. (1987) *Nucleic Acids Res.*, **15**, r53-r188.
33. Varmus,H. and Swanstrom,R. (1984) In Weiss,R., Teich,N., Varmus,H. and Coffin,J.(eds.), RNA Tumor Viruses (2nd edition). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 369–512.
34. Fütterer,J., Gordon,K., Sanfaçon,H., Bonneville,J.-M. and Hohn,T. (1990) *EMBO J.*, **9**, 1697–1707.
35. Proudfoot, N.J. and Brownlee,G.G. (1974) *Nature*, **252**, 359–362.
36. Bouchez,D., Tokuhisa,J.G., Llewellyn,D.J., Dennis,E.S. and Ellis,J.C. (1989) *EMBO J.*, **8**, 4197–4204.
37. Fütterer,J., Gordon,K., Bonneville,J.-M., Sanfaçon,H., Pisan,B., Penswick,J. and Hohn,T. (1988) *Nucleic Acids Res.*, **16**, 8377–8390.
38. Covey,S.N., Turner,D. and Mulder,G. (1983) *Nucleic Acids Res.*, **11**, 251–264.
39. Kozak,M. (1989) *J. Cell Biol.*, **108**, 229–241.
40. Ishikawa,M., Meshi,T., Motoyoshi,F., Takamatsu,N. and Okada,Y. (1986) *Nucleic Acids Res.*, **14**, 8291–8305.
41. Bonneville,J.M., Sanfaçon,H., Fütterer,J. and Hohn,T. (1989) *Cell*, **59**, 1135–1143.
42. Gowda,S., Wu,F.C., Scholthof,H.B. and Shepherd,R.J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9203–9207.
43. Hull,R., Covey,S.N. and Maule,A.J. (1987) *J. Cell. Sci. Suppl.*, **7**, 213–229.
44. Penswick,J., Hübler,R. and Hohn,T. (1988) *J. Virol.*, **62**, 1460–1463.
45. Voytas,D.F. and Ausubel,F.M. (1988) *Nature*, **336**, 242–244.
46. Grandbastien,M.A., Spielmann,A. and Caboche,M. (1989) *Nature*, **337**, 376–380.
47. Lee,D., Ellis,T.H.N., Turner,L., Hellens,R.P. and Cleary,W.G. (1990) *Plant Molec. Biol.* **15**, 707–722.
48. Smyth,D.R., Kalitsis,P., Joseph,J.L. and Sentry,J.W. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 5015–5019.
49. Hull,R., Jones,M.C., Dasgupta,I., Cliffe,J.M., Mingins,C., Lee,G. and Davies,J.W. (1991) *Proceedings 2nd International Rice Genetics Symposium*, International Rice Research Institute, (in press).