# Arenavirus Gene Structure and Organization 

D.H.L. Bishop ${ }^{1}$ and D.D. Auperin ${ }^{2}$

1 Introduction
2 Nucleotide Sequence and Coding Arrangement of the Arenavirus S RNA Species ..... 6
2.1 The Intergenic Region of the Arenavirus S RNA ..... 13
2.2 The $3^{\prime}$ and $5^{\prime}$ Terminal Nucleotide Sequences of Arenavirus L and S RNA Species ..... 15
3 The Arenavirus L RNA ..... 16
References ..... 16

## 1 Introduction

For rhabdoviruses, paramyxoviruses, orthomyxoviruses, and some members of the Bunyaviridae, all the proteins are translated from the viral-complementary RNA sequence. This is in contrast to the genetic strategy of picornaviruses, caliciviruses, coronaviruses, togaviruses, flaviviruses, and retroviruses, whose proteins are all translated from the viral RNA sequence (i.e., viral RNA, or sequences corresponding to viral RNA sequences, functions as an mRNA species). The former group of single-stranded RNA viruses are described as nega-tive-stranded RNA viruses, the latter as positive-stranded RNA viruses. The negative-stranded RNA viruses have a virion RNA polymerase that is responsible for the synthesis of the viral-complementary mRNA species. Although the purified viral RNA of most of the positive-stranded viruses is infectious per se, this is not the case for retroviruses since they have a virion reverse transcriptase that forms an obligatory DNA intermediate during the initial stages of the viral infection process.

Arenaviruses have a genome consisting of two species of single-stranded RNA (Vezza et al. 1978), designated L (large) and S (small). The RNA is not infectious per se (CARTER et al. 1974) and it does not appear to template the synthesis of protein when supplied to competent in vitro translation reactions (Leung et al. 1977). Viral RNA polymerase activities have been identified in Pichinde arenavirus preparations (Carter et al. 1974; Leung et al. 1979). Thus these data indicate that arenaviruses also have a negative-stranded coding strategy. Recent nucleotide sequence analyses of the $S$ genome species of three members of the arenavirus family have shown that the $3^{\prime}$ half of the RNA sequence codes for a protein (the nucleoprotein, NP) in the viral-complementary

[^0]sequence, the $5^{\prime}$ half codes for a protein (the glycoprotein precurser, GPC) in a viral-sense sequence. This strategy is described as an ambisense coding arrangement and is one of the subjects of this chapter.

## 2 Nucleotide Sequence and Coding Arrangement of the Arenavirus S RNA Species

The complete nucleotide sequences of the S RNA species of Pichinde arenavirus, lymphocytic choriomeningitis virus (LCMV) (WE strain) and Lassa fever virus, have recently been determined from partial RNA sequence data (Auperin et al. $1982 \mathrm{a}, \mathrm{b}$ ) and from analyses of representative clones of cDNA (AUPERIN et al. 1984a, b, 1986; Clegg and Oram 1985; Romanowski and Bishop 1985; RomANOWSKI et al. 1985). The results indicate that all three viruses have a similar organization for their S RNA species. The viral and viral-complementary sequences of the S RNA of Pichinde arenavirus are shown in Fig. 1 together with the sequences of the encoded gene products (AUPERIN et al. 1984a, b). The S RNA is reported to be 3419 nucleotides long. The base composition of the viral RNA is $22.2 \% \mathrm{G}, 22.6 \% \mathrm{C}, 26.8 \% \mathrm{~A}$, and $28.4 \% \mathrm{U}$, representing a size of approximately $1.1 \times 10^{6}$ daltons. Similar results have been obtained for LCM and Lassa fever viruses (Clegg and Oram 1985; Romanowski and Bishop 1985; Romanowski et al. 1985; Auperin et al. 1986).

Genetic studies have shown that the arenavirus S RNA species codes for two proteins, NP and GPC (Vezza et al. 1980; Harnish et al. 1983; Riviere et al. 1985). Two methionine-initiated open reading frames that are sufficiently large to encode the two proteins have been identified among the six possible reading frames of the viral and viral-complementary sequences of the S RNA of Pichinde virus (Auperin et al. 1984a, b). One of the reading frames is in the viral-complementary RNA sequence and corresponds to the $3^{\prime}$ half of the viral RNA (Fig. 1 a ). The open reading frame begins at viral-complementary residue 84 and terminates at position 1766. It encodes a protein of 561 amino acids. The protein has been identified as the viral NP based on the observation that antibodies raised to one of its predicted peptides (near the amino terminus of the gene product) precipitate Pichinde NP (Leung et al. 1984). The calculated size ( 62984 daltons), composition, net positive charge ( +9 ), and hydropathic character of the protein (e.g., the absence of hydrophobic domains at either the amino or carboxy termini of the protein) are also consistent with this conclusion (AUPERIN et al. 1984a).

The second S-coded open reading frame is in the $5^{\prime}$ half of the viral RNA sequence (Fig. 1 b; Auperin et al. 1984b). It begins at viral RNA residue 52 and terminates at residue 1560 . The gene product consists of 503 amino acids (size: $57.3 \times 10^{3}$ daltons) and contains two hydrophobic domains, one close to the amino terminus, the other close to the carboxy terminus (Fig. 2). These two regions of hydrophobic amino acids are similar to those of other viral glycoproteins (Rose et al. 1980; Skehel and Waterfield 1975). In addition, the protein is rich in cysteine residues and potential asparagine-linked glycosyla-
tion sites (Fig. 2). Preliminary tryptic peptide sequence data with [ ${ }^{35}$ S]cyteinelabelled tryptic peptides (M. Galinsky and D.H.L. Bishop, unpublished information) indicate that the Pichinde virus GP1 glycoprotein is derived from the amino terminus of GPC and that the GP2 comes from the carboxy terminus. Antibodies made to synthetic peptides from the amino terminal and carboxy terminal regions of the LCMV GPC gene product have recently been shown to react with LCMV GP1 and GP2 glycoproteins, respectively (M. Buchmeier, personal communication).

Shown in Fig. 3 are the flanking nucleotides that surround the NP, GPC, and L RNA translation initiation codons of the Pichinde, LCM, and Lassa fever arenaviruses (Auperin et al. 1985a, b, 1986; Clegg and Oram 1985; Romanowsio and Bishop 1985; Romanowski et al. 1985) compared to those determined by Kozak $(1978,1984)$ to represent the consensus flanking sequence for translation initiation of proteins for eukaryotes (i.e., CCA/GCCAUGG). The data compiled for the indicated translation initiation codons of arenaviruses indicate that there are purines in the -1 and -3 positions (counting the $A U G$ as $+1,+2$, and +3 ) and most of the +4 positions, with a strong preference for an A residue in the -3 position (Fig. 3). These data agree with Kozak's observation that an $A$ or $G$ residue at -3 is frequently found flanking the functional AUG initiation codon. It has been noted (P. Young, personal communication) that in the leader sequence to the Pichinde, LCM, and Lassa fever viruses S and (where it is known) L mRNA species, there is a conserved six nucleotide sequence (GAUCCU, residues 10-15, see Fig. 1) that is complementary to a $3^{\prime}$ terminal sequence (CUAGGA) present in $18 S$ ribosomal RNA (Atmadja et al. 1984). This complementarity may be important in arenavirus mRNA-ribosome interactions and subsequent translation of the mRNA.

The S RNA nucleotide sequence data predict that the mRNA for the arenavirus NP gene must consist of a viral-complementary sequence corresponding to at least the $3^{\prime}$ half of the viral S RNA. Similarly, the mRNA for the GPC gene must correspond to at least the $5^{\prime}$ half of the viral S RNA. Northern blot analyses, using Pichinde or LCM virus-infected cell RNA preparations that were resolved into size classes by gel electrophoresis and hybridized to the appropriate nick-translated viral DNA or terminally labelled, singlestranded, DNA species representing the nucleoprotein genes, have identified subgenomic, viral-complementary, NP mRNA species corresponding to the $3^{\prime}$ halves of the respective S RNA (Auperin et al. 1984b; Romanowski et al. 1985). The subgenomic mRNA has been translated in vitro into NP as identified by the size of the product and by immunoprecipitation analyses. Similar studies using strand-specific probes derived from the glycoprotein gene have detected in infected cell extracts a viral-sense, subgenomic RNA (putative GPC mRNA) corresponding to the $5^{\prime}$ half of the genome (Auperin et al. 1984b). The subgenomic mRNA species present in Pichinde virus-infected cells apparently lack poly A tails on their $3^{\prime}$ termini as evidenced by their inability to bind to oligo dT cellulose columns (Auperin et al. 1984b).

In summary, it has been found that the arenavirus NP is encoded in a viral-complementary, subgenomic mRNA species and that the GPC is encoded in a viral-sense, subgenomic mRNA species. The antiparallel arrangement of








d 1 I S









CCTATTGGGGAGTGCCGTCGAGGCCCATGTCGGAAGCGGAGCTTATTTTCCCAACCTTACCCATTTGTAAGGTTTCTTTGGTATTTATAATACCCGCAGCTACACAGGGAGTTTCTTGTA
ATCCTGTGTGGTtTCGGGCAACCATCACCAATGATGTGCCTATGAGTAGGTATTCCAACTATGTGGAGAAATACTGTAATGGTGTAAAAACCAAAGACCAGAAGCATATATCTGTCAFTT





GTGGTGTTCTGTATGATTAAGAAATTGTAATGTGTTCCACCCTCACAGTTTGTCAGTCTGCAAGTATCTCCACTGCAGTTGTTGAAGCACTTCCCAACCCATGCGATT.TTAGGATCCCCG








10 D.H.L. Bishop and D.D. Auperin

 -

 H ${ }^{\text {T }}$ T




























Fig. 1a, b. Nucleotide sequences of the viral-complementary (a) and viral (b) S RNA species of Pichinde arenavirus presented as the DNA sequence equivalents. The orientation of the sequences are from $5^{\prime}$ to $3^{\prime}$. The amino acid sequence of the gene product (NP) encoded in vc RNA sequence (a, residues 84-1766) is given above the corresponding nucleotide codons. The amino acid sequence of the gene product (GPC) encoded in RNA sequence (b, residues $52-1560$ ) is given above the corresponding nucleotide codons
Fig. 1b

```
Pichinde S vRNA coded GPC
```



Fig. 2. Hydropathic plot, potential glycosylation sites, and charged amino acid distribution of the predicted GPC polypeptide of Pichinde arena-virus. The regions net hydrophobicity (areas above the center line), or net hydrophilicity (are below the center line) are displayed (Kyte and Doolittle 1982) as well as the location of charged amino acids ( $\mathrm{R}, \mathrm{K}, \mathrm{D}, \mathrm{E}$ ) and potential asparagine-linked glycosylation sites ( $Y$, although whether all are employed for glycosylation is not known)

Kozak consensus sequence $C C_{G}^{A}$ CCAUGG

Arenavirus S RNA Viral-complementary RNA:

| Pichinde NP gene | CCAAAAUGU |
| :--- | ---: |
| LCM (WE) NP gene | ACAAAAUGU |
| Lassa NP gene | CGACAAUGA |

Arenavirus L RNA Viral-complementary RNA:

LCM (WE) L gene
GCGCAAUGG
Fig. 3. Comparison of the nucleotides surrounding arenavirus translation initiation codons to those identified by Kozak $(1978,1984)$ (CCA/GCCAUGG) as the most frequent flanking sequence for eukaryotic translation initiation codons


Fig. 4. The intergenic region of Pichinde virus S RNA species. The top sequence shows the viralcomplementary ( $v c R N A$ ) sequence and the carboxy terminal end of the coding region for the N . The bottom sequence shows the viral $(v R N A)$ sequence and the carboxy terminal end of the coding region of the GPC gene product
the two genes on the S RNA is a novel strategy for RNA viruses. The term ambisense RNA has been used to describe it.

### 2.1 The Intergenic Region of the Arenavirus S RNA

The reading frames that code for the NP and GPC genes, as indicated in Fig. 1, do not overlap. For the S RNA of Pichinde virus, there is a short, noncoding, intergenic region of 87 nucleotides between the termination codons of the two genes. The nucleotide sequence within this intergenic region is such that a hairpin structure can be formed that would be stabilized by 14 G-C and 4 A-U base pairs (bp). The structure of the Pichinde virus intergenic region is illustrated in Fig. 4. Similar intergenic regions have been identified for LCM and Lassa fever viruses (Romanowski et al. 1985; Auperin et al. 1986).


Fig. 5. The coding, transcription, and replication strategies of the arenavirus S RNA species

The unique aspect of the ambisense coding strategy is that it provides for the independent regulation of the synthesis of the NP and GPC genes. This is illustrated in Fig. 5. Presumably during a productive infection, transcription of the NP gene occurs immediately after virus adsorption and entry of the viral ribonucleoprotein into the host cell cytoplasm. However, the synthesis of the GPC mRNA cannot occur until viral RNA replication has commenced and full length, replicative intermediate, viral-complementary RNA is available as a template for transcription of the GPC gene. The independent regulation of the expression of these two genes may be important in the establishment and maintenance of persistently infected cells. Gimenez and Compans (1980) have shown that in cells persistently infected with Tacaribe virus the viral GPC is present only at very low levels, although the NP is abundant. Similar observations have been made for Pichinde virus-infected cells (V. Romanowski and D.H.L. Bishop, unpublished data). These observations may be explained if, unlike the NP mRNA synthesis, GPC mRNA synthesis is specifically reduced or inhibited in such cells. One mechanism for selective inhibition could be that there are only small amounts of replicative intermediate, viral-complementary RNA species in those cells, or that such species are not used for GPC mRNA synthesis but are preferentially used for viral RNA replication.

Considering the size of the NP and GPC mRNA species, it is probable that the intergenic region functions as a transcription terminator for both genes. Evidence to support this notion has come from oligonucleotide annealing studies (M. Galinski and D.H.L. Bishop, unpublished data). How during RNA replication the viral polymerase avoids this signal is not known.


Fig. 6. The $3^{\prime}$ terminal nucleotide homologies of the viral and viral-complementary RNA sequences of Pichinde, LCM, and Lassa fever viruses aligned according to their termini (Auperin et al. 1984a, b, 1986; ClegG and Oram 1985; Romanowski and Bishop 1985; Romanowski et al. 1985). The asterisks identify homologous nucleotides. Additional homologies exist if one incorporates gaps in the sequences, in particular for residues 21-35 of the LCMV and Lassa fever virus vRNA and for residues 20-33 of their vcRNA species

### 2.2 The $\mathbf{3}^{\prime}$ and $5^{\prime}$ 'Terminal Nucleotide Sequences of Arenavirus $\mathbf{L}$ and $S$ RNA Species

It has been shown that the $3^{\prime}$ and $5^{\prime}$ termini of the $S$ genome RNA of Pichinde arenavirus are complementary for some 19 nucleotides with two mismatches (Fig. 1, residues 6 and 8; see Auperin et al. 1984b). Similar data have been obtained for the S RNA of LCMV (Romanowski et al. 1985). The complementarity between the terminal nucleotide sequences presumably reflects a consensus sequence required for the initiation of mRNA transcription and RNA replication at the $3^{\prime}$ termini of both the viral and viral-complementary (replicative intermediate) RNA species. It has been reported that there is an extra G nucleotide on the end of the cDNA clones of Pichinde virus corresponding to the $5^{\prime}$ terminus of the S RNA (AUPERIN et al. 1984b). This nucleotide was thought to be a cloning artifact, since its presence interfered with the exact alignment of the complementary termini. A similar observation has been made for Lassa fever virus (Auperin et al. 1986) but was not observed for LCMV (WE strain, Romanowski et al. 1985).

A comparison of the $3^{\prime}$ terminal nucleotide sequences of the viral and viralcomplementary S RNA species of Pichinde, LCMV, and Lassa fever viruses shows that the $3^{\prime}$ terminal 19 nucleotides of the vRNA species are exactly conserved as are the $3^{\prime}$ terminal 19 nucleotides of the vcRNA species (Fig. 6). The three mismatches between the terminal nucleotide sequences of the respective viral and viral-complementary RNA species of each virus and their conservation among the three different arenaviruses suggests that the sequence differences of the $3^{\prime}$ and $5^{\prime}$ ends have a function that has been maintained through evolution. One possible explanation is that the $3^{\prime}$ terminal nucleotide sequence present on the viral and viral-complementary RNA templates may function in determining the efficiency with which mRNA transcription and/or viral RNA replication is initiated from either template. Such a mechanism could provide the basis for the mRNA regulation discussed above.

## 3 The Arenavirus L RNA

Unlike the S RNA, for the L genome RNA species of arenaviruses, there is only a limited amount of nucleotide sequence data available. Romanowski and Bishop (1985) have published the nucleotide sequence for the $3^{\prime}$ terminal 1123 nucleotides of the L RNA of LCMV (WE strain). The data indicate that a viral gene product is encoded in the L viral-complementary RNA sequence beginning at nucleotide residue 33 . The reading frame encoding this gene product remains open for the duration of the sequence determined, encoding a minimum, therefore, of $43 \times 10^{3}$ daltons of protein. Northern blot analyses using the cloned cDNA as a probe identified viral-size RNA species but failed to identify any L subgenomic RNA species. It is presumed that the viral RNA polymerase is encoded on the L RNA. Whether this reading frame encodes other gene products is not known. It is also not known whether the L RNA exhibits an ambisense coding strategy.

## References

Atmadja J, Brimacombe R, Maden BEH (1984) Xenopus laevis 18S ribosomal RNA: experimental determination of secondary structural elements, and locations of methyl groups in the secondary structure model. Nucleic Acids Res 12:2649-2667
Auperin DD, Dimock K, Cash P, Rawls WE, Bishop DHL (1982a) Analyses of the genomes of prototype Pichinde arenavirus and a virulent derivative of Pichinde Munchique: evidence for sequence conservation at the $3^{\prime}$ termini of their viral RNA species. Virology 116:363-367
Auperin DD, Compans RW, Bishop DHL (1982b) Nucleotide sequence conservation at the $3^{\prime}$ termini of the virion RNA species of New World and Old World arenaviruses. Virology 121:200-203
Auperin DD, Galinsky M, Bishop DHL (1984a) The sequences of the N protein gene and intergenic region of the S RNA of Pichinde arenavirus. Virology 134:208-219
Auperin DD, Romanowski V, Galinsky M, Bishop DHL (1984b) Sequencing studies of Pichinde arenavirus S RNA indicate a novel coding strategy, an ambisense viral S RNA. J Virol 52:897-904
Auperin DD, Saso D, McCormick JBM (1986) Nucleotide sequence of the glycoprotein gene and intergenic region of Lassa fever virus. (manuscript in preparation)
Carter MF, Biswal N, Rawls WE (1974) Polymerase activity of Pichinde arenavirus. J Virol 13:577-583
Clegg JCS, Oram JD (1985) Molecular cloning of Lassa virus RNA: nucleotide sequence and expression of the nucleoprotein gene. Virology 144:363-372
Gimenez HB, Compans RW (1980) Defective interfering Tacaribe virus and persistent infected cells. Virology 107:229-239
Harnish DG, Dimock K, Bishop DHL, Rawls WE (1983) Gene mapping in Pichinde virus: assignment of viral polypeptides to genomic $L$ and S RNAs. J Virol 46:638-641
Kozak M (1978) How do eucaryotic ribosomes select initiation regions in messenger RNAs. Cell 15:1109-1123
Kozak M (1984) Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature 308:241-246
Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132
Leung W-C, Ghosh HP, Rawls WE (1977) Strandedness of Pichinde virus RNA. J Virol 22:235-237
Leung W-C, Leung MFKL, Rawls WE (1979) Distinctive RNA transcriptase, polyadenylic acid polymerase, and polyuridylic acid polymerase activities associated with Pichinde virus. J Virol 30:98-107

Leung W-C, Ramsingh A, Jing G, Mong K, Taneja AK, Hodges RS (1984) Subgenomic RNA of arenavirus Pichinde and its implication in regulation of viral gene expression in productive infection and persistence. Academic, New York
Riviere Y, Ahmed R, Southern PJ, Buchmeier MJ, Dutko FJ, Oldstone MBA (1985) The S RNA segment of lymphocytic choriomeningitis virus codes for the nucleoprotein and glycoproteins 1 and 2. J Virol 53:966-968
Romanowski V, Bishop DHL (1985) Conserved sequences and coding of two strains of lymphocytic choriomeningitis virus (WE and ARM) and Pichinde arenavirus. Virus Res 2:35-51
Romanowski V, Matsura Y, Bishop DHL (1985) Complete sequence of the S RNA of lymphocytic choriomeningitis virus (WE strain) compared to that of Pichinde arenavirus. Virus Res 3:101-114
Rose JK, Welch WJ, Sefton BM, Esch FS, Ling NC (1980) Vesicular stomatitis virus glycoprotein is anchored in the viral membrane by a hydrophobic domain near the COOH terminus. Proc Natl Acad Sci USA 77:3884-3888
Skehel JJ, Waterfield MD (1975) Studies of the primary structure of the influenza virus hemagglutinin. Proc Natl Acad Sci USA 72:93-97
Vezza AC, Clewley JP, Gard GP, Abraham NZ, Compans RW, Bishop DHL (1978) Virion RNA species of the arenaviruses Pichinde, Tacaribe, and Tamiami. J Virol 26:485-497
Vezza AC, Cash P, Jahrling P, Eddy G, Bishop DHL (1980) Arenavirus recombination: the formation of recombinants between prototype Pichinde and Pichinde Munchique viruses and evidence that arenavirus S RNA codes for N polypeptide. Virology 106:250-260


[^0]:    ${ }^{1}$ NERC Institute of Virology, Mansfield Road, Oxford OX1 3SR, United Kingdom
    ${ }^{2}$ Centers for Disease Control, Atlanta, GA 30333, USA

