Journal of General Virology (1997), **78**, 1571–1576. Printed in Great Britain

The complete Mokola virus genome sequence: structure of the RNA-dependent RNA polymerase

P. Le Mercier, Y. Jacob and N. Tordo

Laboratoire des Lyssavirus, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France

The genome sequence of the rabies-related virus Mokola virus (genus *Lyssavirus*) has been completed by sequencing the L gene, which consists of 6384 nucleotides encoding a 2127 amino acid polymerase. Alignment of the Mokola virus L protein with other polymerases from the virus order *Mononegavirales* defined three domains: a divergent NH_2-terminal domain, a highly conserved central domain carrying most of the functional motifs and a COOH-terminal domain with alternating conserved and divergent regions. A statistical study outlined the stringency of conservation of glycine, acidic (D, E) and basic (K, R, H) amino acids in polymerases, particularly as key residues of the conserved motifs.

Mokola virus is a rabies-related virus that forms genotype 3 of the genus *Lyssavirus* (family *Rhabdoviridae*) (Bourhy *et al.*, 1993). It is among the most genetically distant from rabies virus (genotype 1), and modern anti-rabies vaccines fail to protect against Mokola virus infection. Mokola virus has been responsible for a few cases of human and animal encephalomyelitis, largely dispersed throughout sub-saharian Africa (Foggin, 1982). The induced pathology is similar to that of rabies, although no typical 'furious' form has been associated with Mokola virus infection (King *et al.*, 1994). In mice, Mokola virus is less virulent than rabies virus since it does not kill when injected by a peripheral route (Perrin *et al.*, 1996). *In vitro*, Mokola virus can multiply in mosquito (*Aedes albopictus*) cells whereas rabies virus cannot (Aitken *et al.*, 1984). This characteristic is shared only by very distant lyssaviruses such as Obodhiang and Kotokan, which are true arboviruses (Buckley, 1975). Indeed, Mokola virus has been isolated from insectivorous mammals such as shrews (Shope *et al.*, 1970).

The Mokola virus genome is a single negative-stranded RNA molecule (order *Mononegavirales*) which encodes five proteins. Two interact with the viral membrane: the matrix protein (M or M2), which is involved in virion morphology, and the transmembrane glycoprotein (G), which provides cell receptor recognition, and against which neutralizing antibodies are directed. The other three proteins in the ribonucleocapsid structure participate in transcription and replication: the nucleoprotein (N) tightly encapsidates the genome, making it a template for the polymerase complex which comprises the cofactor phosphoprotein (P or M1) and the RNA-dependent RNA polymerase (L). The latter is a nucleotidyl transferase (RNA synthesis), a guanylyl transferase (capping) and a poly(A) synthetase (polyadenylation). In several mononegavirales, L protein has also been shown to be a specific kinase for the phosphoprotein, although the functional role of this phosphorylation remains to be demonstrated (Gao & Lenard, 1995).

We have completed the sequence of the Mokola virus genome (EMBL accession number Y09762) by sequencing the complete L gene using three overlapping cDNA clones – *β5β* (3295 bp), pM12 (2815 bp) and R15*α* (710 bp) – previously described (Bourhy *et al.*, 1989, 1993). Inserts were subcloned into the pUC18 vector, serially deleted by restriction enzyme or exonuclease III digestion, and then sequenced on both strands using either M13 forward or reverse primer and automated sequencing (ALF Pharmacia).

The Mokola virus genome is the third lyssavirus genome to be completely sequenced (Tordo *et al.*, 1988; Conzelmann *et al.*, 1990). Its 11939 nucleotides (nt) are extensively used for coding purposes since 90·3% (10782 nt) code for the N, P, M, G and L proteins, and only 7·3% (941 nt) are neither protein- nor known signal-coding sequences. This further underlines the importance of the large Ψ non-coding region (504 nt) preceding the L gene, which is found in all lyssavirus genomes but whose function remains unknown (Tordo *et al.*, 1986). The similarity profile obtained by comparison of the genomes of Mokola virus and rabies virus strain PV (Fig. 1A) shows, as expected, that the non-coding regions are less well-conserved than the coding regions. Of the coding regions, the N gene is the most conserved (81·7% amino acid identity), followed, in decreasing order, by the L (77·9%), M (76·3%), G (58·4%) and P (47·4%) genes. The L open reading frame (ORF), 6384 nt in length, is flanked by two sequences homologous to the consensus transcription start (genome position 5416) and stop (11862) signals previously characterized (Bourhy *et al.*, 1989). Fig. 1(B) compares the non-translated sequences of rhabdovirus

Author for correspondence: Noël Tordo.
Fax +33 1 40 61 32 56. e-mail ntordo@pasteur.fr

The nucleotide sequence reported in this paper will appear in the EMBL nucleotide sequence database under accession number Y09762.

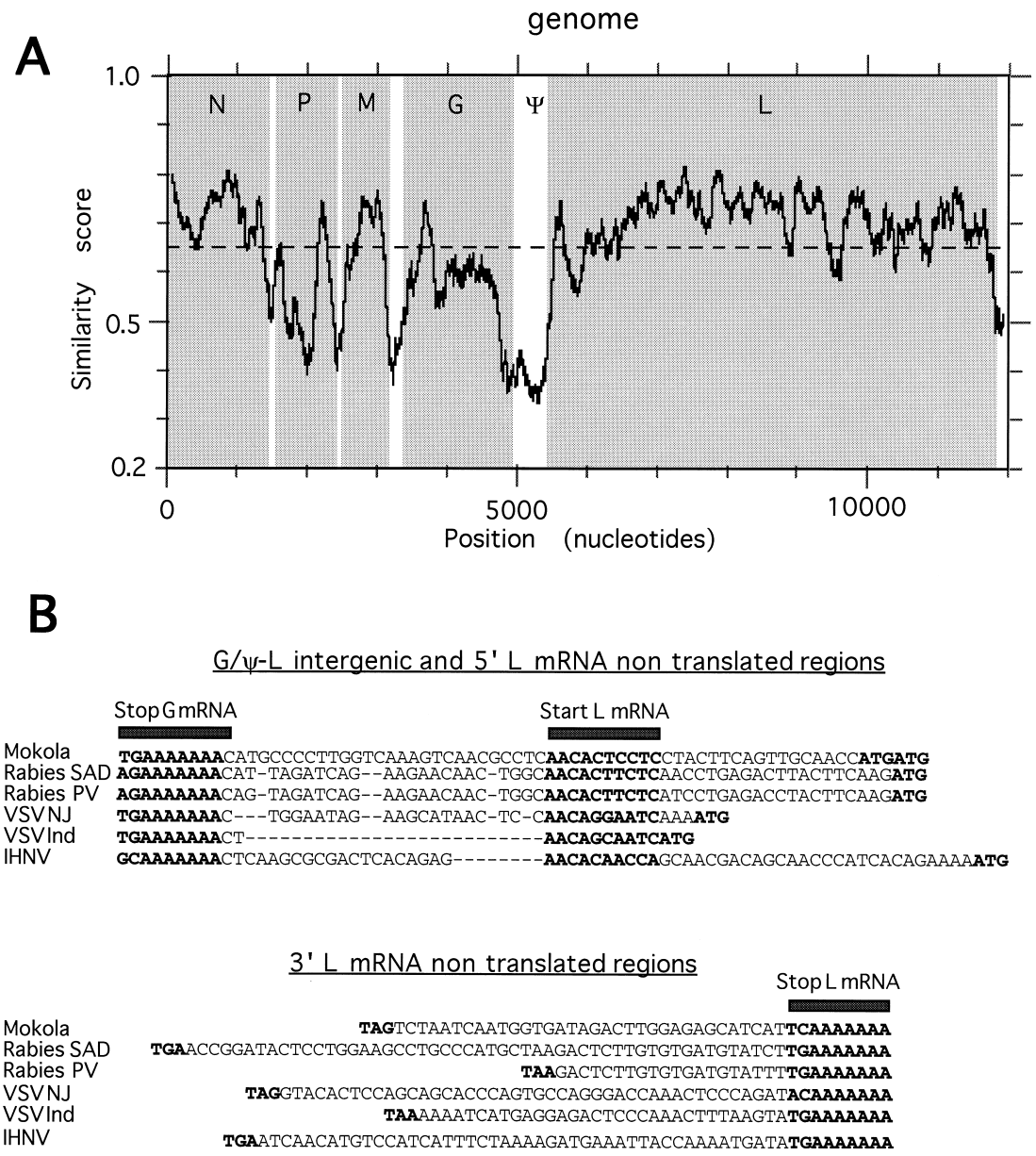

Fig. 1. (A) Plotsimilarity profile (GCG, window 150) between rabies PV and Mokola virus genomes. (B) Alignment of the G–L intergenic region and the 5′ and 3′ non-translated regions of rhabdovirus L mRNAs.

L mRNAs. The 5′ ends, although very short, are larger in lyssaviruses and infectious haematopoietic necrosis virus (IHNV) (30–35 nt) than in vesicular stomatitis virus (VSV) (10–15 nt). The Mokola virus ORF typically starts with two adjacent AUG codons from position 5443 : the second of these, which is present only in rabies virus strains, is in a better context for translation initiation and extends up to a UAA codon at position 11824–26. Interestingly, the lyssavirus L proteins show heterogeneity at their COOH termini. In rabies virus SAD strain the ORF stops six codons before that of Mokola virus which in turn is five codons shorter than in rabies virus PV strain. The untranslated 3′ ends of the rhabdovirus L mRNAs are 30–60 nt in length. Finally, the untranscribed intergenic region between the G mRNA stop and the L mRNA start is rather long (20–28 nt) in lyssaviruses, IHNV and VSV New Jersey, but very short (2 nt) for VSV Indiana. This intergenic flexibility could play a role in the regulation of transcription.

The deduced size (2126–2127) and amino acid composition of the Mokola virus L protein are very similar to those of the genotype 1 counterparts : rabies virus strains PV (Tordo *et al.*, 1988) and SAD-B19 (Conzelmann *et al.*, 1990). Such relatedness opens new perspectives for the delineation of the conserved domains and key functional motifs essential for polymerase function. A previous comparison of mononegavirales L proteins provided evidence that conservation was not equally

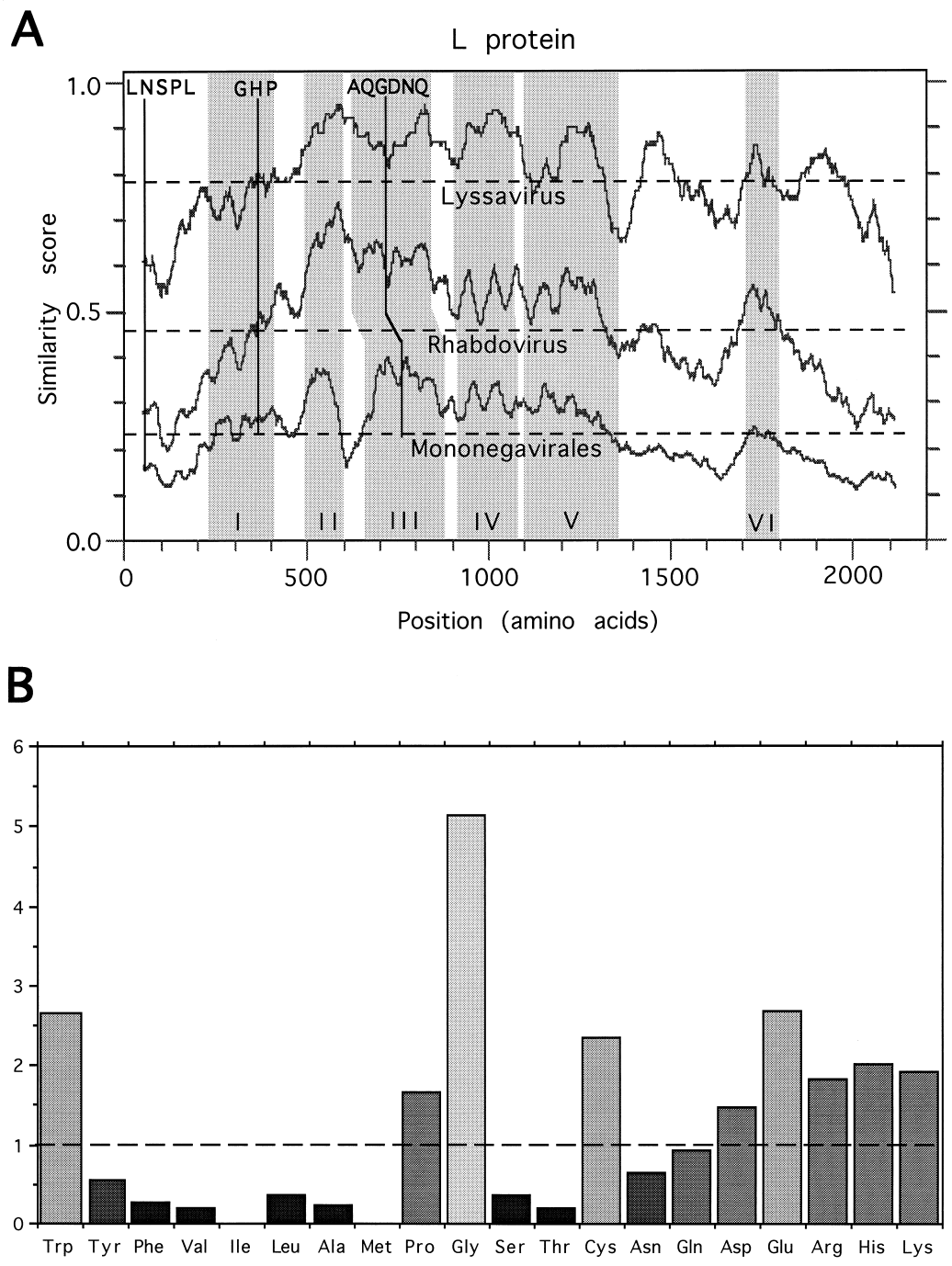

Fig. 2. (A) Plotsimilarity profile (GCG, window 100) between lyssavirus (rabies PV and Mokola virus), rhabdovirus (lyssaviruses plus VSV Indiana and New Jersey) and mononegavirales (rhabdoviruses plus measles virus Udem, Newcastle disease virus and Sendai virus strain Z) polymerases, showing the six conserved blocks and the principal motifs discussed in the text.
(B) Relative presence of each amino acid among the 97 invariant residues found when 12 polymerases, representative of the main phylogenetic groups of mononegavirales (indicated with asterisks in the phylogenetic tree in Fig. 3) were compared. For each amino acid, the value shown corresponds to its frequency in the invariant positions divided by its overall frequency in the L protein.

distributed throughout the L gene and delineated six conserved blocks (I–VI) (Tordo *et al.,* 1988; Poch *et al.,* 1990). Fig. 2(A) shows the plotsimilarity profiles when comparing the L proteins of two lyssaviruses, four rhabdoviruses or seven mononegavirales. Despite the decrease in average score from 78% (lyssaviruses) to 47% (rhabdoviruses) and 23% (mono-

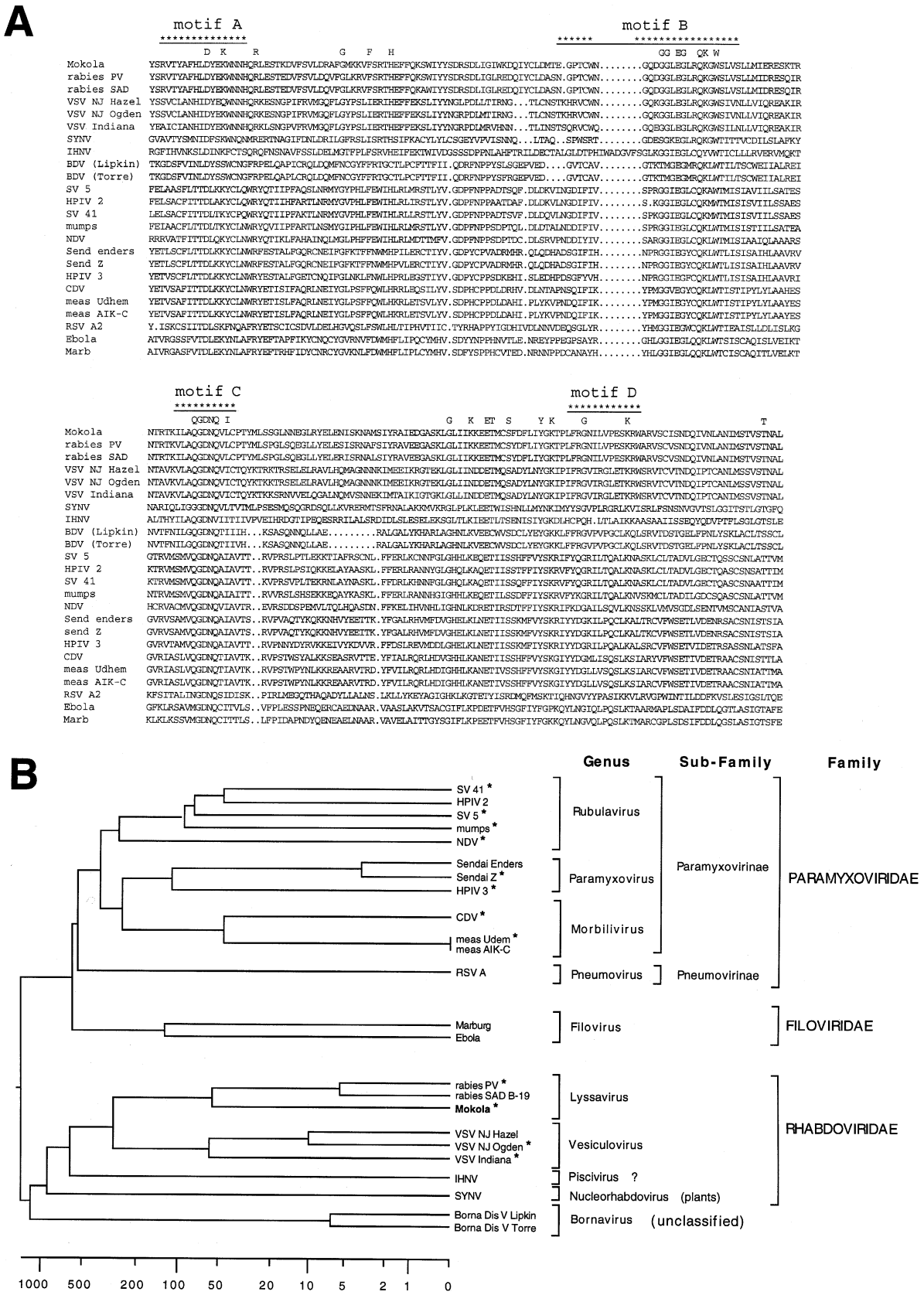

Fig. 3. For legend see facing page.

negavirales), the overall profile remains similar and distinguishes three principal domains linked by hinge regions. (1) A divergent NH_2-terminal domain from which block I emerges, although rather poorly, when distant proteins are compared (mononegavirales). However this domain includes strongly invariant motifs, like the 'LNSPL' motif (position 38), close to the NH_2 terminus (Poch *et al.*, 1990). (2) A large central 'polymerase' domain containing blocks II–V. This domain is the most conserved although the block delineation varies slightly between profiles. In particular, a variable interblock, region II–III, in the mononegavirales profile is longer in paramyxovirus than in rhabdovirus polymerases, and requires the introduction of a gap. (3) A large COOH-terminal domain, including block VI, where conservation is intermittent and on average lower. Interestingly, this domain comprises three conserved blocks in the lyssavirus profile, but becomes very divergent in the rhabdovirus and mononegavirales profiles, where only the central block VI is rather poorly maintained. Such a region, showing conservation within a virus genus but divergence between virus families, is most likely to encode genus-specific functions or to be indicative of specific interactions with variable genes or proteins. In this context, one should note that the binding of the L polymerase with its cofactor, the variable P protein, involves the COOH-terminal domain in VSV (Canter & Perrault, 1996) and the divergent NH_2-terminal domain in measles virus (Horikami *et al.*, 1994).

Enzymatic activities have been predicted for several conserved blocks and/or motifs on the basis of L protein sequence comparisons (Poch *et al.*, 1989, 1990). However, their functional evaluation was only recently possible using the newly developed reverse genetic tools (Schnell *et al.*, 1994; Conzelmann, 1996). Block III of the central 'polymerase' domain, particularly the core motif AQGDNQ (motif C; see Fig. 3A) has been intensively dissected by introduction of mutations or deletions. This motif was predicted to be essential for RNA polymerase function because it is similar in all RNA-dependent polymerases (Poch *et al.*, 1989). Investigation of both rabies virus (Schnell & Conzelmann, 1995) and VSV (Sleat & Banerjee, 1992) L proteins confirmed that almost any change made to the motif resulted in complete loss of polymerase activity. Furthermore, assessment of the relative importance of the flanking residues showed that the only change tolerated was asparagine (N) to aspartate (D), corresponding to a change of the 'DN' signature of mononegavirales polymerases into the 'DD' signature of reverse transcriptases and polymerases of positive-stranded RNA viruses (Poch *et al.*, 1989).

In the hinge region (nt 1360–1520) separating the central and the COOH-terminal domains, a phenylalanine (F-1488) to serine (S) mutation has been implicated in thermosensitivity of VSV mutants with an aberrant polyadenylation phenotype (Hunt & Hutchinson, 1993). In an attempt to reproduce this phenotype in the Sendai virus L protein, mutations were generated at the analogous residue [cysteine (C)-1571] (Horikami & Moyer, 1995). They were either neutral, or inhibited or enhanced the polymerase function, or inhibited transcription without altering replication. Similarly, by mimicking the measles virus L protein around the conserved GHP motif of block I (position 372) (Chandrika *et al.*, 1995), it was hoped to give the Sendai virus L protein the ability to bind the measles virus P protein (Horikami *et al.*, 1994). However, uncoupling between transcription and replication was obtained. These results show the limitations of predictive analysis, and illustrate the gap between sequence comparisons and knowledge of the three-dimensional structure of the active protein. They also emphasize that very divergent regions of the L protein may play an important functional role, either per se (NH_2 domain) or by promoting cooperation between domains (hinge region between central and COOH-terminal domains).

In order to assess whether specific amino acids in the polymerase sequence are subject to particular selective pressure for conservation, 12 L proteins representative of the mononegavirales as a whole were aligned. The frequency of each amino acid in the invariant population versus its frequency in the entire population was calculated. A value > 1 signifies a tendency to conservation; conversely, a value < 1 means poor conservation. Fig. 2(B) shows that apart from tryptophan (W), which classically is highly conserved, glycine (G), cysteine (C), proline (P), acidic [aspartate (D), glutamate (E)] and basic [arginine (R), lysine (K), histidine (H)] amino acids are more conserved and typically play a key role in L protein function and/or structure. On the other hand, serine (S), threonine (T), hydrophobic [alanine (A), isoleucine (I), leucine (L), valine (V), methionine (M)] and aromatic [tyrosine (Y), phenylalanine (F)] amino acids are very poorly conserved. Asparagine (N) and glutamine (Q) are rather neutral. Stringent conservation of G and of acidic and basic amino acids was previously inferred from their frequent involvement in strongly invariant motifs (Poch *et al.*, 1989, 1990). It was also verified by site-directed mutagenesis in block III described above: mutations of G and charged residues led to loss of function, while hydrophobic residues could be substituted for each other without affecting activity (Schnell & Conzelmann, 1995). Such observations imply that the matrix of penalties attributed to each mismatch during polymerase alignment should be adjusted accordingly. Indeed, the matrix used by default in most alignment programs is statistically calculated based on changes observed between

Fig. 3. (A) Alignment of the conserved block III of 24 L proteins of mononegavirales. Amino acids conserved in at least 20 out of the 24 sequences are noted above the alignment. Motifs A to C were initially described in all RNA-dependent polymerases (Poch *et al.*, 1989). Note that they match well with the conserved residues although mononegavirales show some typical features (an insertion within motif B because of IHNV; a larger motif D). (B) Phylogenetic tree constructed by the neighbour-joining method (Clustal W) and based on the alignment in (A). Asterisks indicate the viruses selected for the analysis in Fig. 2(B).

proteins in the databanks, and is not typical of a functional L polymerase.

The high conservation of L proteins separated by large genetic distances points to them as the best targets for phylogenetic studies. For example, the L and N proteins are almost equally conserved among lyssaviruses (77·9% vs 81·7%), but only L retains significant similarity among mononegavirales. Using the neighbour-joining method, the amino acid sequences of the conserved block III of 24 L proteins were compared (Fig. 3 A). The phylogenetic lineages obtained largely respect the taxonomic cleavages of the order Mononegavirales (Fig. 3 B): the families *Rhabdoviridae*, *Paramyxoviridae*, *Filoviridae* and the unclassified Borna disease virus are distinct. Subclassification into genera is also observed. Interestingly, the rate of divergence is significantly smaller among lyssavirus (22·1% between rabies and Mokola virus) than vesiculovirus (34·8% between VSV Indiana and New Jersey) L proteins. This suggests that lyssaviruses could more easily exchange proteins without altering functionality. In contrast, exchange of proteins involved in transcription and replication is almost impossible between VSV New Jersey and Indiana (Smallwood *et al*., 1994). This opportunity for genetic exchange in the genus *Lyssavirus* opens interesting perspectives for flexibility in developing a reverse genetic approach using recombinant technology. Functional investigation of the blocks and motifs of the L protein as well as analysis of the relative independence of the NH_2-, central and COOH-domains is envisaged.

The authors thank Dr Hervé Bourhy for his contribution to cDNA cloning, and Dr Katherine Kean for critical review of this manuscript. P.L.M. was a recipient of a C.A.N.A.M. fellowship.

References

Aitken, T. H. G., Kowalski, R. W., Beaty, B. J., Buckley, S. M., Wright, J. D., Shope, R. E. & Miller, B. R. (1984). Arthropod studies with rabies-related Mokola virus. *American Journal of Tropical Medicine and Hygiene* **33**, 945–952.

Bourhy, H., Tordo, N., Lafon, M. & Sureau, P. (1989). Complete cloning and molecular organization of a rabies-related virus: Mokola virus. *Journal of General Virology* **70**, 2063–2074.

Bourhy, H., Kissi, B. & Tordo, N. (1993). Molecular diversity of the Lyssavirus genus. *Virology* **194**, 70–81.

Buckley, S. M. (1975). Arbovirus infection of vertebrate and insect cell cultures, with special emphasis on Mokola, Obodhiang and Kotonkan viruses of the rabies serogroup. *Annals of the New York Academy of Sciences* **266**, 241–250.

Canter, D. M. & Perrault, J. (1996). Stabilization of vesicular stomatitis virus L polymerase protein by P protein binding: a small deletion in the C-terminal domain of L abrogates binding. *Virology* **219**, 376–386.

Chandrika, R., Sandra, M., Horikami, S. M., Smallwood, S. & Moyer, S. A. (1995). Mutations in conserved domain I of the Sendai virus L polymerase protein uncouple transcription and replication. *Virology* **213**, 352–353.

Conzelmann, K. K. (1996). Genetic manipulation of non-segmented negative-strand RNA viruses. *Journal of General Virology* **77**, 381–389.

Conzelmann, K. K., Cox, J. H., Schneider, L. G. & Thiel, H. J. (1990). Molecular cloning and complete nucleotide sequence of the attenuated rabies virus SAD B19. *Virology* **175**, 485–489.

Foggin, C. M. (1982). Atypical rabies virus in cats and a dog in Zimbabwe. *Veterinary Record* **110**, 338.

Gao, Y. & Lenard, J. (1995). Multimerization and transcriptional activation of the phosphoprotein (P) of vesicular stomatitis virus by casein kinase-II. *EMBO Journal* **14**, 1240–1247.

Horikami, S. M. & Moyer, S. A. (1995). Alternative amino acids at a single site in the Sendai virus L protein produce multiple defects in RNA synthesis *in vitro*. *Virology* **211**, 577–582.

Horikami, S. M., Smallwood, S., Bankamp, B. & Moyer, S. A. (1994). An amino proximal domain of the L protein binds to the P protein in the measles virus RNA polymerase complex. *Virology* **205**, 540–545.

Hunt, D. M. & Hutchinson, K. L. (1993). Amino acid changes in the L polymerase protein of vesicular stomatitis virus which confer aberrant polyadenylation and temperature-sensitive phenotypes. *Virology* **193**, 786–793.

King, A. A., Meredith, C. D. & Thomson, G. R. (1994). The biology of Southern African Lyssavirus variants. *Current Topics in Microbiology and Immunology* **187**, 267–295.

Perrin, P., De Franco, M., Gontier-Jallet, C., Fouque, F., Morgeaux, S., Tordo, N. & Colle, J. (1996). Antigen-specific cell-mediated immune response in mice is suppressed by infection with pathogenic lyssaviruses. *Research in Virology* **147**, 289–299.

Poch, O., Sauvaget, I., Delarue, M. & Tordo, N. (1989). Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO Journal* **8**, 3867–3874.

Poch, O., Blumberg, B. M., Bougueleret, L. & Tordo, N. (1990). Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignments of functional domains. *Journal of General Virology* **71**, 1153–1162.

Schnell, M. & Conzelmann, K. K. (1995). Polymerase activity of *in vitro* mutated rabies L protein. *Virology* **214**, 522–530.

Schnell, M. J., Mebatsion, T. & Conzelmann, K. K. (1994). Infectious rabies viruses from cloned cDNA. *EMBO Journal* **13**, 4195–4203.

Shope, R. E., Murphy, F. A., Harrison, A. K., Causey, O. R., Kemp, G. E., Simpson, D. I. H. & Moore, D. L. (1970). Two African viruses serologically and morphologically related to rabies virus. *Journal of Virology* **6**, 690–692.

Sleat, D. E. & Banerjee, A. K. (1992). Transcriptional activity and mutational analysis of recombinant vesicular stomatitis virus RNA polymerase. *Journal of Virology* **67**, 1334–1339.

Smallwood, S., Sumners, E. R. & Moyer, S. A. (1994). Determinants of serotype specificity in transcription of vesicular stomatitis virus synthetic nucleocapsids. *Virology* **199**, 11–19.

Tordo, N., Poch, O., Ermine, A., Keith, G. & Rougeon, F. (1986). Walking along the rabies genome: is the large G–L intergenic region a remnant gene? *Proceedings of the National Academy of Sciences, USA* **83**, 3914–3918.

Tordo, N., Poch, O., Ermine, A., Keith, G. & Rougeon, F. (1988). Completion of the rabies virus genome sequence determination: highly conserved domains along the L (polymerase) proteins of unsegmented negative-strand RNA viruses. *Virology* **165**, 565–576.

Received 11 December 1996; Accepted 14 March 1997