Available online at www.sciencedirect.com
SCIENCE DIRECT®
Virology 318 (2004) 153–168
VIROLOGY
www.elsevier.com/locate/yviro

Sequence analysis of L RNA of Lassa virus☆

Simon Vieth,[a] Andrew E. Torda,[b] Marcel Asper,[a] Herbert Schmitz,[a] and Stephan Günther[a,*]

[a] *Department of Virology, Bernhard-Nocht-Institute for Tropical Medicine, 20359 Hamburg, Germany*
[b] *Zentrum für Bioinformatik, University of Hamburg, 20146 Hamburg, Germany*

Received 13 June 2003; returned to author for revision 23 July 2003; accepted 12 September 2003

Abstract

The L RNA of three Lassa virus strains originating from Nigeria, Ghana/Ivory Coast, and Sierra Leone was sequenced and the data subjected to structure predictions and phylogenetic analyses. The L gene products had 2218–2221 residues, diverged by 18% at the amino acid level, and contained several conserved regions. Only one region of 504 residues (positions 1043–1546) could be assigned a function, namely that of an RNA polymerase. Secondary structure predictions suggest that this domain is very similar to RNA-dependent RNA polymerases of known structure encoded by plus-strand RNA viruses, permitting a model to be built. Outside the polymerase region, there is little structural data, except for regions of strong alpha-helical content and probably a coiled-coil domain at the N terminus. No evidence for reassortment or recombination during Lassa virus evolution was found. The secondary structure-assisted alignment of the RNA polymerase region permitted a reliable reconstruction of the phylogeny of all negative-strand RNA viruses, indicating that *Arenaviridae* are most closely related to *Nairoviruses*. In conclusion, the data provide a basis for structural and functional characterization of the Lassa virus L protein and reveal new insights into the phylogeny of negative-strand RNA viruses.

Keywords: Lassa virus; Arenavirus; Phylogeny; Sequence; Secondary structure prediction; RNA-dependent RNA polymerase

Introduction

Lassa virus belongs to the family *Arenaviridae*. Its natural host is the African rodent *Mastomys natalensis*. Transmission of the virus to humans causes Lassa fever which is associated with hemorrhage and organ failure and has a high case fatality rate among hospitalized patients. Lassa fever is endemic in the West African countries of Sierra Leone, Guinea, Liberia, and Nigeria. A case of Lassa fever in a traveller who visited Ivory Coast, Ghana, and Burkina Faso during the incubation period (Günther et al., 2000) suggests that the virus is also prevalent outside the previously established endemic areas of West Africa.

Arenaviridae, which comprise only the genus arenaviruses, are classified as segmented negative-strand RNA viruses together with the families *Bunyaviridae* and *Orthomyxoviridae*. Arenaviruses are divided phylogenetically, serologically, and geographically into two major complexes, the Old World complex (e.g. Lassa virus and lymphocytic choriomeningitis virus [LCMV]) and the New World complex (e.g. Tacaribe virus and Pichinde virus) (Bowen et al., 1997).

The single-stranded arenavirus genome consists of a small (S) and a large (L) RNA segment, 3.4 and 7 kb, respectively, in length. Arenaviruses use a so-called ambisense coding strategy. Two genes are located on each RNA segment in opposite directions separated by an intergenic region. This region is predicted to have a very stable secondary structure. The 3′- and 5′-terminal 19 nucleotides of the RNA segments are complementary to each other and are highly conserved among all arenaviruses. Comprehensive sequence information on the S RNA of several arenaviruses, including Lassa virus, has been accumulated in recent years (Bowen et al., 1997, 2000; Charrel et al., 2002; Garcia et al., 2000; Weaver et al., 2000) and full-length S RNA sequences of Lassa virus are available for all major phylogenetic lineages: for strains from Sierra Leone (Auperin and McCormick, 1989), from the region comprising Ivory Coast, Ghana, and Burkina Faso (Günther et al., 2000), and from Nigeria (Bowen et al., 2000; Clegg et al.,

☆ Lassa virus RNA sequences have been sent to GenBank and assigned the accession numbers AY179171–AY179175. Supplementary data for this article are available on ScienceDirect.

* Corresponding author. Department of Virology, Bernhard-Nocht-Institute for Tropical Medicine, Bernhard-Nocht-Strasse 74, D-20359 Hamburg, Germany. Fax: +49-40-42818-378.
E-mail address: guenther@bni.uni-hamburg.de (S. Günther).

0042-6822/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2003.09.009

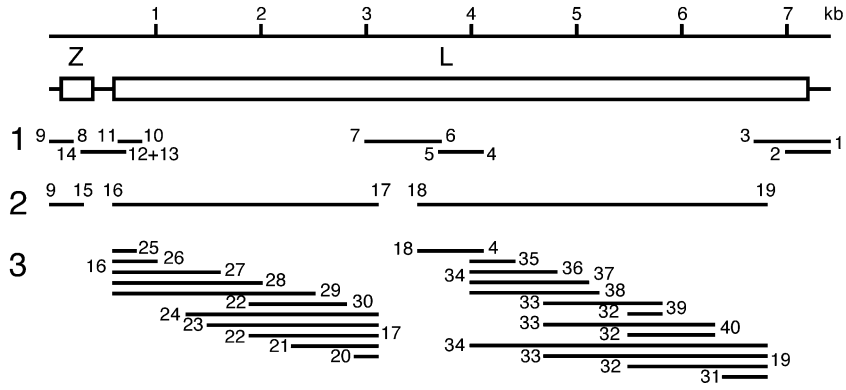


Fig. 1. Three-step strategy for sequencing Lassa virus L RNA. Firstly, primers were designed on the basis of known L RNA sequences of Lassa virus Josiah and LCMV and used for RT-PCR. Secondly, the fragments were sequenced, strain-specific primers were designed, and used for long-range RT-PCR. Thirdly, shorter fragments were amplified under low stringency conditions using the long-range fragments as a template and were directly sequenced. The gene organization of L RNA and a scale bar is shown at the top. Horizontal lines represent PCR products obtained with Lassa virus CSF as an example. The corresponding primers are indicated by numbers. The primer sequences are listed according to these numbers in the supplementary data for this article.

1991; Günther et al., 2001). Phylogenetic analysis of partial as well as full-length S RNA sequences indicates a geographical pattern (from east to west) in Lassa virus evolution within West Africa (Bowen et al., 2000). In contrast, sequence information on the L RNA segment is much more limited. So far, only the L RNA sequences of Lassa virus Josiah (Djavani et al., 1997; Lukashevich et al., 1997), LCMV WE and Armstrong (Djavani et al., 1998; Salvato and Shimomaye, 1989; Salvato et al., 1989; Singh et al., 1987), Tacaribe virus (Iapalucci et al., 1989a,b), and Pichinde virus are known.

The S RNA encodes the glycoprotein precursor (GPC) and the nucleoprotein (NP) while the L RNA encodes the 250-kDa L protein and the 11-kDa Z protein. The Z protein contains a zinc-binding RING domain and is a structural component of the virion (Salvato et al., 1992). Due to its interaction with a variety of cellular proteins (Borden et al., 1998a,b; Campbell Dwyer et al., 2000), it is also assumed to have a regulatory role. The L protein most likely contains the RNA-dependent RNA polymerase that is associated with the viral nucleocapsid (Fuller-Pace and Southern, 1989; Singh et al., 1987). Direct evidence for an essential role of L protein in genome transcription and replication was recently provided with arenavirus replicon systems (Lee et al., 2000; Lopez et al., 2001). However, more extensive studies on structure and function of the L protein are needed.

Table 1
General features of Lassa virus L RNA

Lassa virus		L RNA				Reference
Strain	Origin	5′ NCR (nt)	Z gene (nt)	L gene (nt)	3′ NCR (nt)	
Josiah	Sierra Leone	65	300	6657	157	(Djavani et al., 1997; Lukashevich et al., 1997)
NL	Sierra Leone	68	300	6663	158	this article
AV	Ghana, Côte D'Ivoire, or Burkina Faso	66	300	6663	137	this article
CSF	Nigeria	70	300	6654	77	this article

Table 2
Identity at the nucleotide and amino acid level among arenaviruses

Gene	% Identity							
	Lassa Josiah vs. Lassa NL		Among Lassa lineages		Lassa vs. LCMV		Lassa vs. New World viruses	
	nt	aa	nt	aa	nt	aa	nt	aa
NP	93.6	96.6	78.8	92.2	61.7	63.9	56.2	51.1
GPC	93.2	97.5	80.6	92.9	61.8	60.7	49.8	42.3
Z	93.3	93.9	74.1	78.0	53.3	55.3	46.7	39.7
L	92.5	95.8	74.3	81.5	53.3	48.1	45.7	35.8

Abbreviations: nt, nucleotide level; aa, amino acid level.

Viral RNA-dependent polymerases share conserved sequences in their catalytic domain, called motifs pre-A [sometimes referred to as motif F (Lesburg et al., 1999)], A, B, C, D, and E (Müller et al., 1994; Poch et al., 1989). Since these enzymes are encoded by many viruses, their sequences have been used to infer the phylogenetic relationship above the virus family level (Bruenn, 1991; Koonin, 1991; Koonin and Dolja, 1993; Lukashevich et al., 1997; Marriott and Nuttall, 1996). However, reevaluation of some data sets indicated that they actually do not contain sufficient information to reconstruct phylogenetic relationships between most virus taxa (Zanotto et al., 1996). Phylogenies recently reconstructed for segmented negative-strand RNA viruses (*Bunya*-, *Orthomyxo*-, and *Arenaviridae*) differed remarkably and the relationships between a number of taxa could not be resolved (Lukashevich et al., 1997; Marriott and Nuttall, 1996).

The crystal structures of RNA-dependent RNA polymerases have recently been determined from three single-stranded, positive-strand RNA viruses of the family *Flaviviridae* (hepatitis C virus [HCV]) (Ago et al., 1999; Bressanelli et al., 1999; Lesburg et al., 1999), *Picornaviridae* (polio virus) (Hansen et al., 1997), and *Caliciviridae* (rabbit hemorrhagic disease virus [RHDV]) (Ng et al., 2002). All three RNA polymerases share a common structure in spite of high sequence divergence (Ng et al., 2002).

In the present study, the L and Z genes of three Lassa virus strains of different geographic origin were sequenced. The sequence data were used for phylogenetic analyses and structure predictions taking advantage of known structures of related RNA-dependent RNA polymerases. The study provides a basis for biochemical and genetic analysis of the Lassa virus L protein and provides new insights into the phylogeny of negative-strand RNA viruses.

Results

Sequencing of L RNA

A three-step strategy was chosen to sequence the L RNA segments of Lassa virus strains NL, AV, and CSF (Fig. 1). Since Lassa viruses are highly diverse in sequence, primers were initially designed in regions showing some degree of conservation among the known L RNA sequences of Lassa virus Josiah and LCMV Armstrong and WE. Depending on

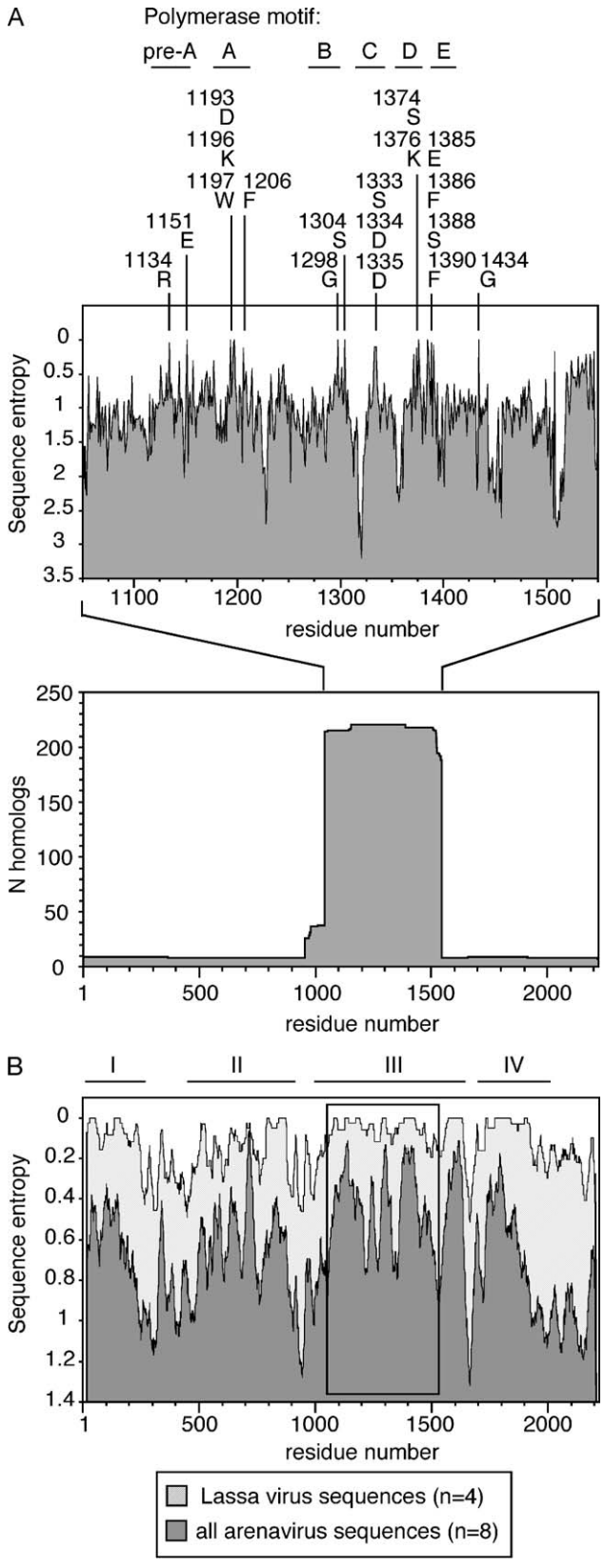


Fig. 2. Sequence homologies across the Lassa virus L protein. (A) Number of homologues of Lassa virus L protein found by two phase, iterated psi-BLAST searches. The number of sequence homologues at each site was counted, considering only those with an expectation value less than 1×10^{-100}. The upper part shows the sequence entropy of the putative RNA polymerase region. The calculations used a conventional 20-residue alphabet. The 220 reliable homologues with an expectation value less than 1×10^{-100} were used in the calculation. The amino acid residues at the highly conserved sites (entropy close to 0) and the conserved polymerase motifs are indicated above the plot. (B) Sequence entropy for L protein of Lassa virus and related arenaviruses. The entropy calculation used a six-residue-type alphabet and a sliding window of 31 residues was used for smoothing. Regions with clustering of conserved sites are marked above the plot.

the strain, a variable number of short regions could be reverse transcribed and amplified using these primers. The fragments were sequenced to design strain-specific primers for long-range RT-PCR. The long-range PCR products were used as a template for amplification of a number of short fragments using internal primers under low stringency conditions. These fragments were sequenced using the PCR primers. Remaining gaps were closed after designing new sets of strain-specific primers.

During PCR of the intergenic region, which predictably contains a highly stable stem-loop structure, products accumulated that were shorter than expected suggesting a deletion in the fragment. Addition of PCR enhancer solution for GC-rich sequences led to an enrichment of the correct fragment which was gel-purified and used for sequencing. However, upon sequencing, the stem-loop region was again missing, indicating that deletions were also generated during the sequencing reaction. The correct sequence was eventually obtained for Lassa virus AV and NL after applying additives and higher reaction temperatures. The full-length stem-loop sequence of strain CSF remained refractory to sequencing despite the use of different thermostable polymerases for sequencing.

Sequence comparison

The Lassa virus L genes ranged in length from 6654 to 6663 nucleotides, while all Z genes, including that of Lassa virus Josiah, had a length of 300 nucleotides (Table 1). The noncoding regions at the 3′ end were generally longer than those at the 5′ end. Both showed a high degree of length and sequence heterogeneity. Some conserved nucleotides in the noncoding regions were found downstream of the conserved RNA termini as well as at positions which form the KOZAK sequence immediately upstream of the start codon (see supplementary data for sequence alignments). The predicted RNA stem-loop structure in the intergenic region was completely conserved (see supplementary data).

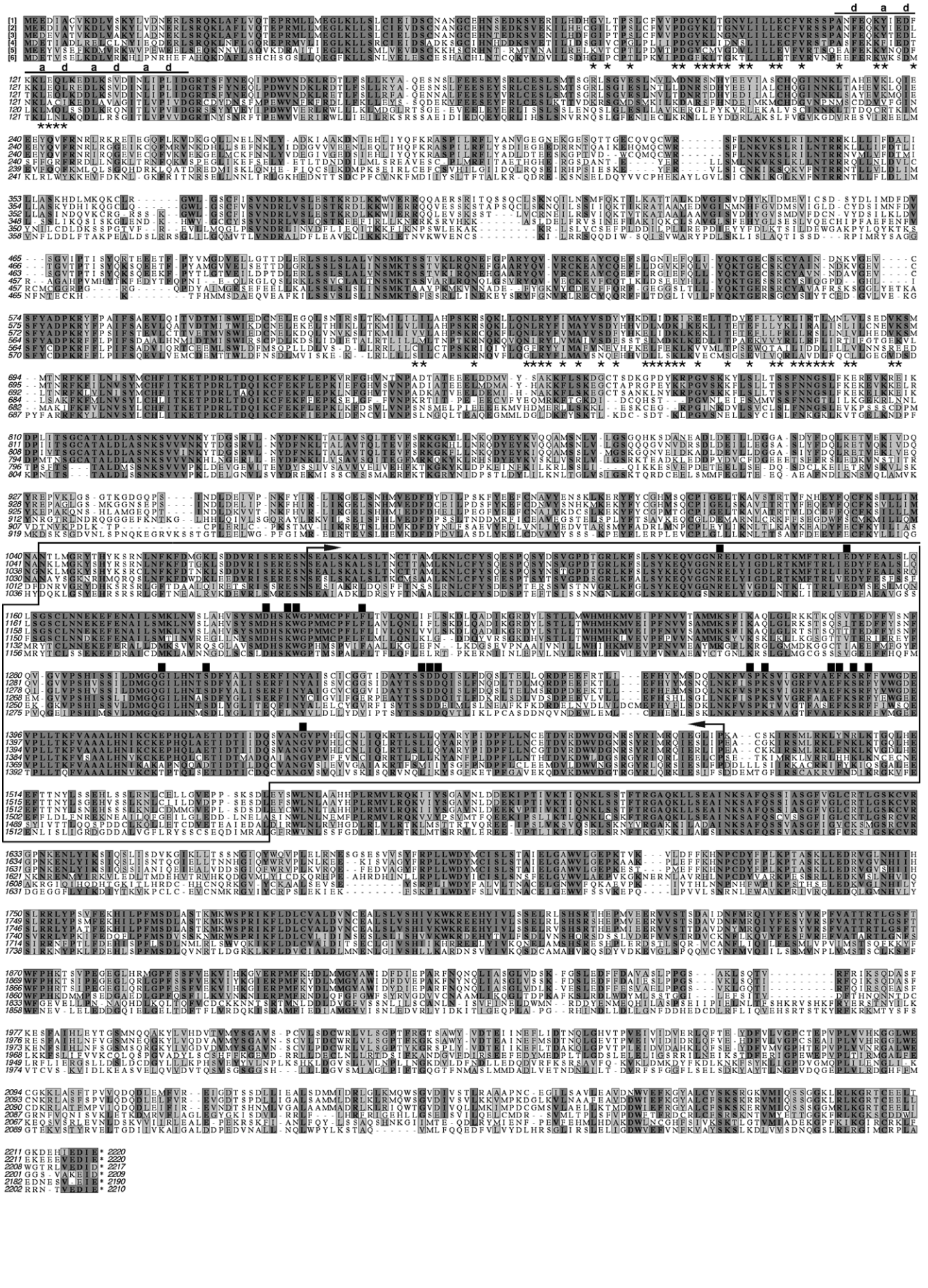
To determine the overall degree of sequence variability among Lassa virus strains as well as between Lassa virus and other arenaviruses, nucleotide and amino acid sequences of each gene were aligned and compared (Table 2). The sequences of the two Lassa virus strains from the same geographic region (Josiah and NL) were very similar. However, when Lassa virus strains of different geographic regions were compared, both L and Z genes differed by about 26% at the nucleotide level and about 20% at the amino acid level, which was considerably higher than among the structural genes GPC and NP. Similarly, when Lassa virus was compared with other Old World and New World arenaviruses, L and Z genes were generally less conserved than GPC and NP genes (Table 2).

The database search using the Lassa virus L protein sequence revealed to what extent the L protein is similar to any known sequence. Fig. 2A (bottom) shows a simple count of the number of aligned residues at each position from sequences with a final expectation value less than 1×10^{-100} in the two phase, iterated psi-BLAST searches. Essentially, for most of the sequence (N and C terminus), the only homologues were from the other arenaviruses. In contrast, the region from residues 1043 to 1546 shows quite a clear border for a putative RNA-dependent RNA polymerase domain. For this entire region, there are more than 200 sequences (or fragments). Inspection of the homologues shows that they are entirely viral RNA polymerases of segmented negative-strand RNA viruses. Dominating the list are PB1 sequences from influenza virus strains (n = 184), followed by L protein sequences of bunyaviruses (n = 23) and tenuiviruses (n = 5). If one considers the default acceptance criterion (expectation <0.005), there are 340 sequences, of which 284 came from *Orthomyxoviridae*, 36 from *Bunyaviridae*, and 5 from tenuiviruses. Statistically, this many homologues produces a very reliable profile for database searching.

The conventional quantity of sequence entropy was used to assess variation within the homologues. This measure is really only meaningful when a large number of sequences are considered, so the analysis was initially confined to the region containing the putative RNA polymerase. Fig. 2A (top) shows the sequence entropy here. There are sites where the entropy is near to zero; these residues are most conserved among the polymerase sequences of all segmented negative-strand RNA viruses. The conserved residues were nearly exclusively located within the known polymerase sequence motifs (Fig. 2A, top and Fig. 3; D1193, D1334, and D1335 are the putative catalytic residues).

To assess sequence variability across the whole L protein, the sequence entropy was calculated in a modified way. Because there are so few reliable homologues—which is insufficient for statistics on 20 amino acids—a reduced alphabet was used as described in Materials and methods (aliphatic, aromatic, polar, positive, negative, and special). Even then, the statistical significance of conserved sites would be dubious, so it was prudent to use a sliding window. The chosen window size of 31 residues is arbitrary, but is much smaller than a protein domain, while large enough to remove much of the noise. Fig. 2B shows this analysis for Lassa virus only (four sequences) and for all

Fig. 3. Comparison of L protein sequences of Lassa virus NL [1], AV [2], CSF [3], LCMV WE [4], Pichinde virus [5], and Tacaribe virus [6]. Shaded residues represent the most conserved residues within the Lassa virus L protein. The putative RNA polymerase domain is boxed and residues conserved among the segmented negative-strand RNA viruses are marked by black rectangles above the sequence (see Fig. 2A). Arrows encompass the region for which the structure was predicted (Fig. 4). Stars below the sequence indicate positions outside the polymerase domain which are conserved between the L protein of arenaviruses and hantaviruses (Hantaan virus L protein positions 85–138 and 610–678). The predicted coiled-coil domain in the Lassa virus L protein is marked with a bar above the sequence. The hydrophobic "a" and "d" positions are indicated.



arenaviruses (eight sequences). There is evidence of sequence conservation in four regions of the L protein: at the N terminus up to position 250 (region I); between position 500 and 900 (region II); between position 1000 and 1650 (region III); and between position 1750 and 1900 (region IV). These regions contain sequences which are virtually completely conserved among all arenaviruses (Fig. 3). As expected, the putative RNA polymerase domain in region III shows the largest number of conserved sites (Fig. 3, boxed). Conserved regions are interspersed by stretches of high sequence entropy (i.e. variability) around positions 300, 950, and 1700. The C terminus from position 1900 is the least conserved part of the protein (Fig. 2B).

A low-stringency iterated psi-BLAST search confined to the sub-database of virus sequences (expectation value 0.05 in the second phase) identified some homologies at a statistically meaningless level between arena- and hantavirus L protein sequences in the conserved regions I and II (Fig. 3, marked by asterisks). These regions correspond to those described previously (Müller et al., 1994).

The Z protein was conserved only within the RING finger motif and at the very N and C termini (see supplementary data for an amino acid sequence alignment).

Structure predictions

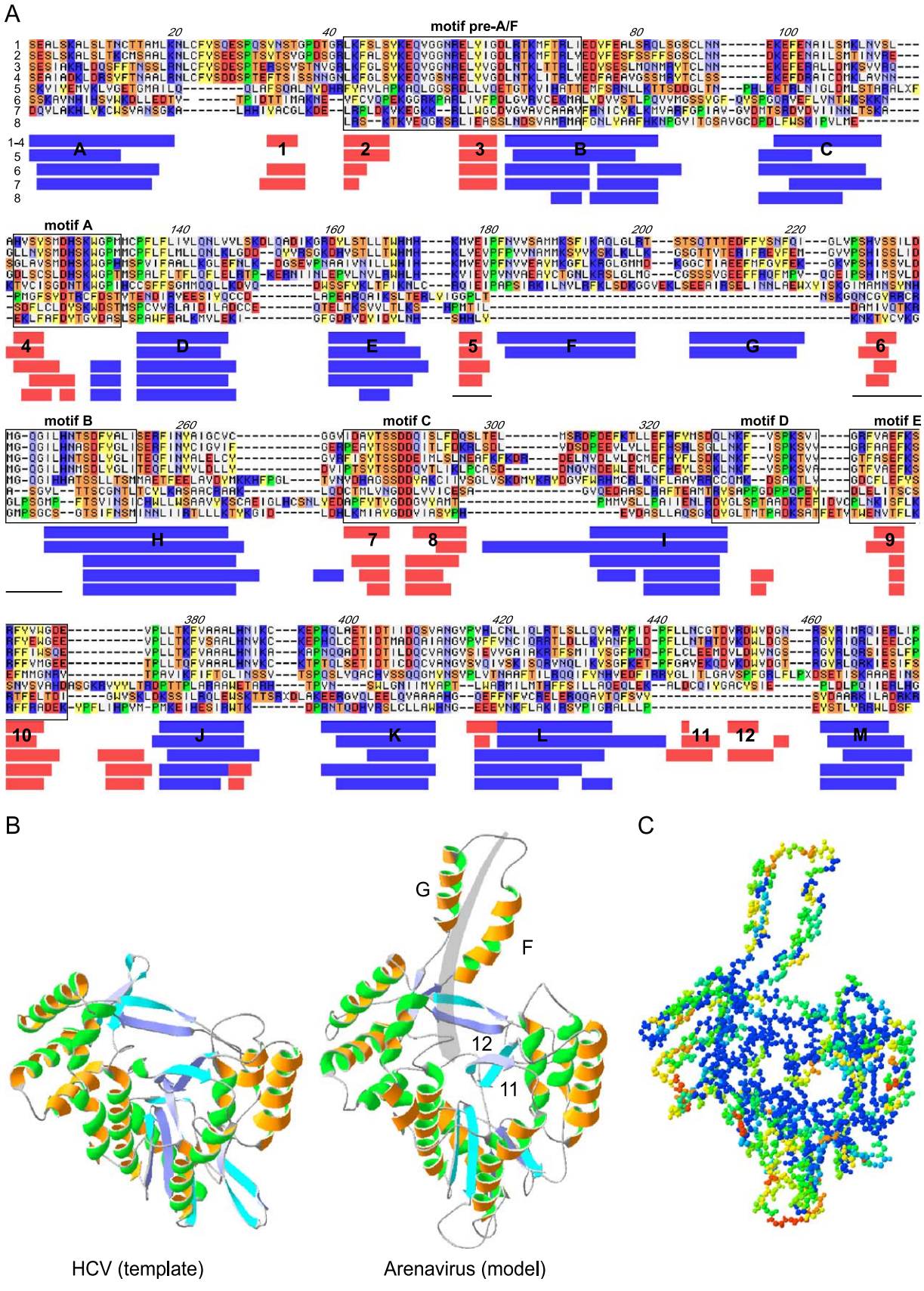
The sequence analysis did not suggest any biological function outside of the putative RNA polymerase region, even with a generous statistical cut-off. One might then hope that a sequence threading approach would detect structural homologues, in spite of the absence of sequence similarity. However, there were no predictions of statistical significance with overlapping L protein sequence fragments ranging from 150 to 500 residues with either the PSSM server based on Boltzmann statistics (Bates et al., 2001) or the code based on a *z*-score optimised scoring function (Huber et al., 1999). More surprisingly, neither program detected similarity to any RNA polymerase of known structure in the region where the sequence-based methods did show a clear signal.

In the case of the locally run threading code, the results were analyzed in more detail to see if there are any patterns. In the N-terminal region, between residues 100 and 350, there was a series of hits to alpha-helical proteins such as 1HCI (alpha-aktinin, 72% alpha-helical, coiled coil) and 1I49 (arfaptin 2; 78% alpha-helical) and further downstream between residues 500 and 700 to proteins such as 1CUN (alpha spectrin; 85% helical). Given the lack of confidence, these cannot be called real structural homologues, but the results are suggestive of alpha-helical domains in the N terminus.

These results led to a search for coiled-coil domains using a specialised server (Lupas et al., 1991). The first calculations suggested a coiled-coil domain at positions 110–138 with a probability of 0.8–1.0 in all Lassa virus sequences, but not in the L protein sequence of LCMV, Pichinde virus, and Tacaribe virus. When the hydrophobic positions of the coiled coil (a and d) were given a higher weight in the analysis, the probability was lower. This resulted from the presence of a single hydrophilic residue (aspartate 119) at a d position (Fig. 3). If this residue was manually set to a hydrophobic one (leucine), the prediction was for a continuos coiled-coil domain with probability 1.0, suggesting that the aspartate could be a so-called buried hydrophilic residue as found in real coiled-coil structures (Akey et al., 2001).

The structures of the RNA polymerases of HCV, polio virus, and RHDV (Ago et al., 1999; Bressanelli et al., 1999; Hansen et al., 1997; Lesburg et al., 1999; Ng et al., 2002) share the same fold and a virtually identical arrangement of secondary structure elements despite low or undetectable sequence homology. We therefore wondered whether the putative RNA polymerase domain of arenaviruses might share the secondary structure arrangement and thus possibly also the overall folding with these three polymerases. As first step in the analysis, the crystal structures of the RNA polymerases of HCV, RHDV, and polio virus (1C2P, 1RDR, and 1KHV, respectively) were superimposed and homologous secondary structure elements as well as amino acid positions were determined. This purely structural alignment served as a seed for the complete alignment. Next, the secondary structures of the putative polymerase domain of Lassa virus Josiah, NL, AV, and CSF; LCMV Armstrong and WE; Pichinde virus; and Tacaribe virus were predicted using the neuronal network program Jnet, which reaches a prediction accuracy of 76% (Cuff and Barton, 2000). To i-

Fig. 4. Secondary structure and folding model of the putative polymerase domain of the arenavirus L protein. (A) Alignment of amino acid sequences and secondary structure elements (below) of Lassa virus CSF [1], LCMV WE [2], Pichinde virus [3], Tacaribe virus [4], Dugbe virus [5], HCV [6], RHDV [7], and polio virus [8]. The partial alignment shown in the figure rests upon an alignment that includes a total of 33 negative-strand RNA viruses. A graphical presentation of a complete alignment is provided in the supplementary data. The secondary structures of negative-strand RNA viruses were predicted, while those of HCV, RHDV, and polio virus are X-ray crystallographic data (Ago et al., 1999; Bressanelli et al., 1999; Hansen et al., 1997; Lesburg et al., 1999; Ng et al., 2002). Alpha helices (blue bars) and beta strands (red bars) are labelled by letters and numbers, respectively. Areas which are distorted in the crystal structure of polio virus are indicated by a line. Conserved sequence motifs are boxed. Chemically similar amino acid residues were colour-coded as follows: acidic (D and E), red; basic (R, K, and H), blue; hydroxyl group (S and T), orange; aromatic (F, Y, and W), yellow; amido group (N and Q), light blue; proline (P), green; aliphatic and sulphur group (L, I, V, G, A, M, and C), light gray. In the sequence of Dugbe virus, deletions of 12–14 amino acids (X) were introduced into loop regions to save space. (B) Folding model of the arenavirus polymerase using the central part of the HCV polymerase as a template (typical front view into the catalytic site). Helices F and G, which have no homologous structure in the template, were modelled by extending the loop present at that position in the template. The possible position of the template strand according to previous modelling studies (Bressanelli et al., 1999) is indicated by a grey arrow. (C) The backbone of the model was coloured according to sequence diversity among all eight arenavirus sequences from dark blue (completely conserved) to red (highly variable) (same view as in B).


A
motif pre-A/F
motif A
motif B
motif C
motif D
motif E
B
C
G
F
12
11
HCV (template)
Arenavirus (model)

ase the accuracy of the analysis, the predicted secondary structures and amino acid sequences of the RNA polymerase region of 25 related viruses were included in the alignment [*Bunyaviridae*/tenuivirus (n = 6), *Orthomyxoviridae* (n = 4), *Paramyxoviridae* (n = 6), *Filoviridae* (n = 2), *Rhabdoviridae* (n = 6), and *Bornaviridae* (n = 1)]. Finally, secondary structure elements and corresponding amino acid sequences were aligned, guided by conserved sequence motifs pre-A, A, B, C, D, and E.

With minor variations, identical elements were predicted in all negative-strand viruses (Fig. 4A, and data no shown). The majority of predicted elements could be clearly assigned to homologous elements of the crystal structures. A few differences are remarkable. First, a helical region (helices F and G) was predicted between beta strands 5 and 6 in all segmented negative-strand viruses (*Arena*-, *Bunya*-, and *Orthomyxoviridae*). This region is missing in the crystal structures as well as in the non-segmented negative-strand RNA viruses which were predicted to have an extended loop at that position. Second, arenavirus polymerases may contain a beta-loop structure (strands 11 and 12) between helices L and M, a structural element which was observed so far only in the HCV polymerase and has been shown to be functionally important in HCV replication (Cheney et al., 2002; Hong et al., 2001; Zhong et al., 2000). Third, a beta strand present in the crystal structures (between strand 10 and helix J) was missing in arena- and bunyaviruses, but was predicted to be present in the orthomyxoviruses and non-segmented negative-strand RNA viruses.

Based on the secondary structure alignment, a hypothetical model of the arenavirus polymerase was generated using the structure of the HCV polymerase as a template (Fig. 4B). The model shows a possible position of the additional putative helices F and G and the beta strands 11 and 12 within the overall structure. Given the low degree of sequence homology between model and template, the quality of the model was not sufficient to predict interactions at the atomic level. When the polypeptide backbone of the model was coloured according to sequence variability among the eight arenavirus sequences (Fig. 4C), the central part of the structure, that is, the catalytic site, was found to be highly conserved, while peripheral parts of the molecule, often predicted loop sequences, were variable. Thus, the position of conserved and variable residues within the sequence is in agreement with the proposed model.

Phylogenetic analysis

Phylogenetic relationships among the arenavirus species were analyzed using full-length NP, GPC, L, and Z genes (Fig. 5). The analysis included Lassa virus Josiah, NL, AV, and CSF; LCMV Armstrong and WE; Pichinde virus; and Tacaribe virus. Both the maximum likelihood (ML) and neighbor-joining (NJ) method of PHYLIP (Felsenstein, 1995) inferred the same phylogeny for all genes with high bootstrap support values at most branches. Thus, no evidence was found for reassortment between L and S RNA segments or recombination of entire genes during evolution

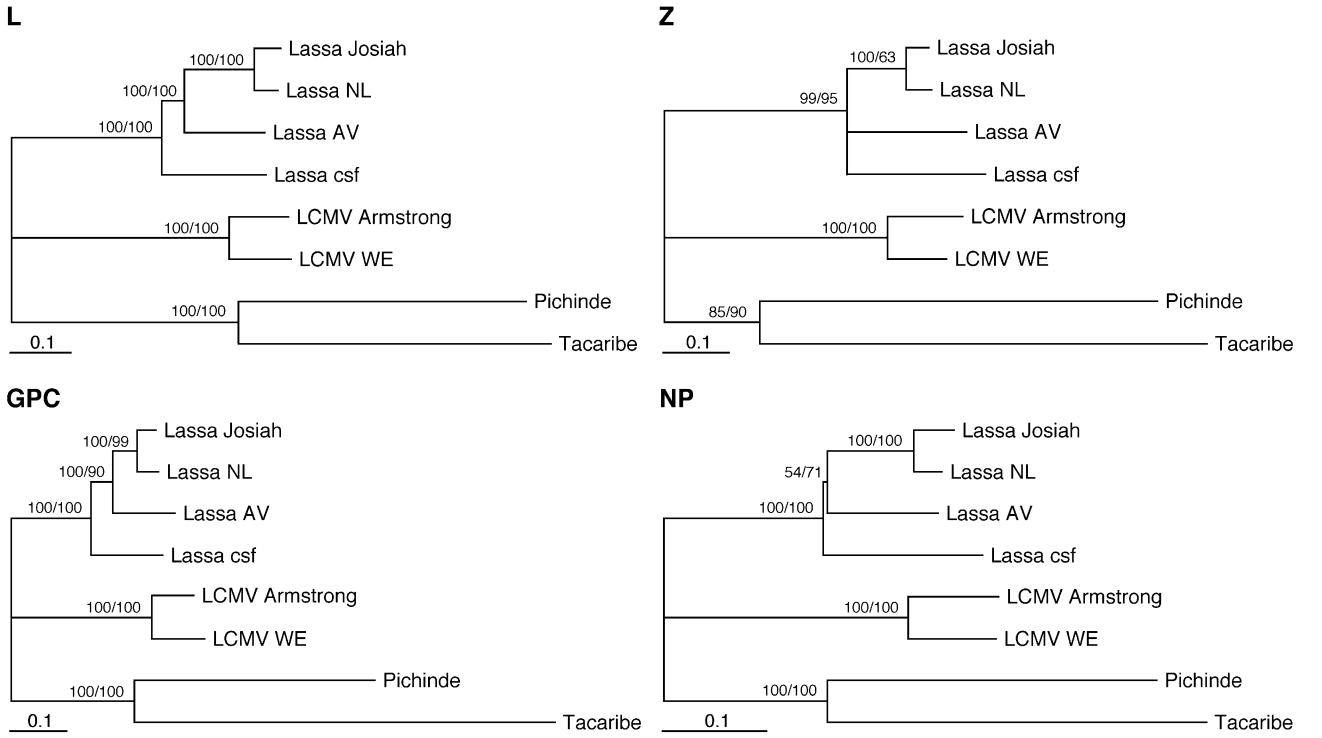


Fig. 5. Phylogenetic relationships among arenaviruses analyzed with full-length L, Z, GPC, and NP genes. Trees generated by NJ (shown in the figure) and ML method were identical. Bootstrap support values are indicated at the respective branches (NJ/ML).

of Lassa virus. There was also no evidence that recombination took place within any of the genes as analyzed with the substitution distribution-based GENECONV test (Sawyer, 1999) and the phylogeny-based RDP test (Martin and Rybicki, 2000), both of which are among the most sensitive tests for detecting recombination (Posada and Crandall, 2001).

Finally, we were interested in reconstructing the phylogenetic relationships above the virus family level, in particular between the *Arenaviridae* and other segmented and non-segmented negative-strand RNA viruses. A major problem that arises in inferring phylogenies of distantly related sequences is the correct assignment of homologous positions in the multiple alignment. We presumed that the

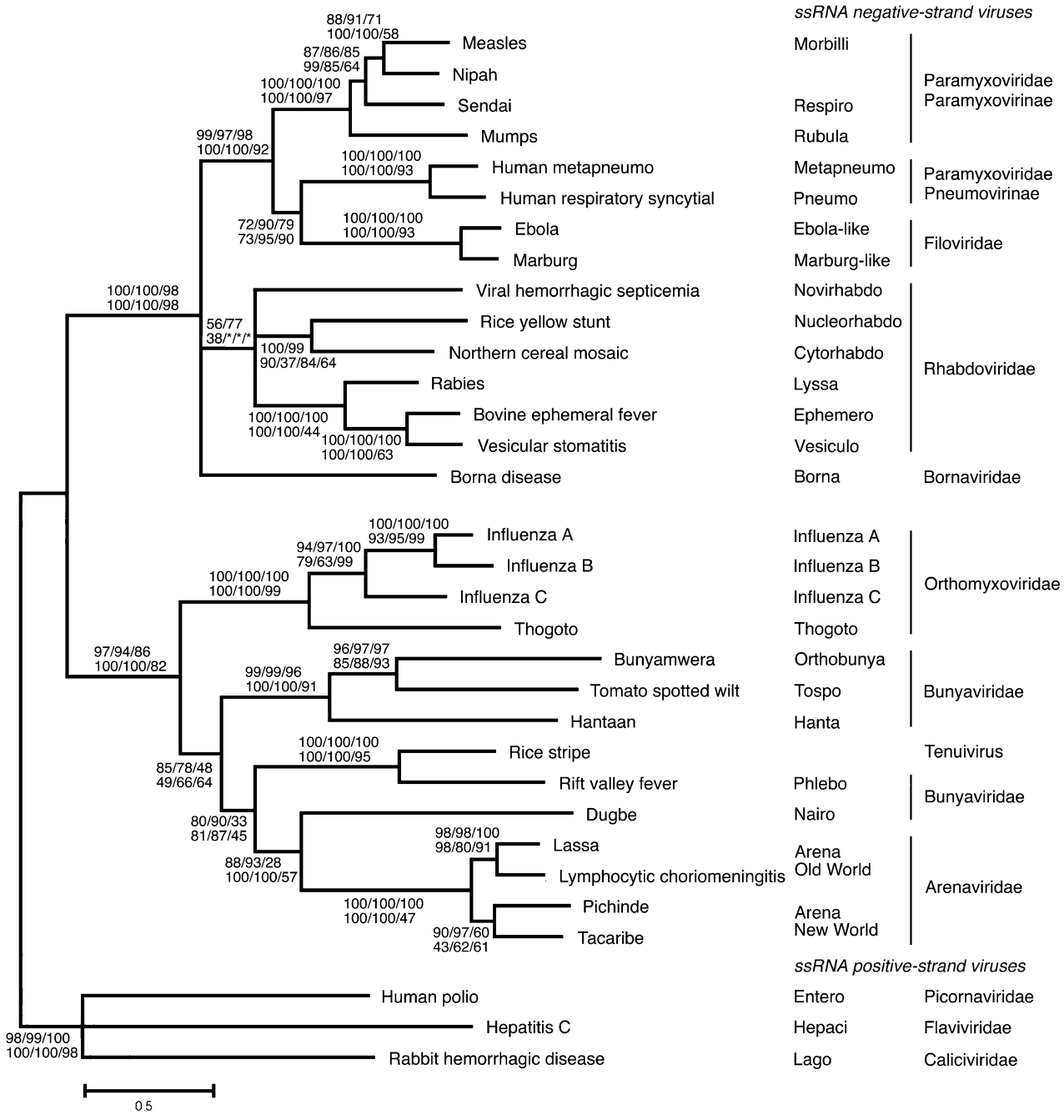


Fig. 6. Inference of phylogenetic relationships among negative-strand RNA viruses using the secondary structure-assisted alignment of polymerase sequences (shown in the supplementary data for this article). Phylogenies were reconstructed by NJ (shown in the figure), ME, and parsimony algorithms as well as by Bayesian probability analysis and ML analysis using quartet-puzzling. Branches were consistently inferred by all methods, except of the basal branches of the nonsegmented negative-strand viruses (collapsed or marked with *). Bootstrap and credibility values are indicated at the branches in the following order: NJ/ME/parsimony, and beneath Bayesian (JONES matrix)/Bayesian (DAYHOFF matrix)/ML quartet-puzzling. The taxonomic nomenclature (genera and families) is indicated on the right.

accuracy of the alignment, especially in regions which lack obvious sequence similarity, can be increased if the sequences are initially aligned on the basis of structural data. For this reason, the secondary structure-assisted alignment of the polymerase sequences (see above) was chosen as a basis for the phylogenetic analysis. The full-length alignment was trimmed by removing most of the highly variable intervening loop sequences, leaving over mainly sites in common secondary structure elements, namely in alpha helices B, C, D, E, H, I, J, K, and L, and beta strands 2, 3, 4, 5, 6, 7, 8, 9, and 10 (numbering according to Fig. 4A). These sites were assembled into an alignment of 313 amino acids in length (see supplementary data for the complete alignment). Though the accuracy of an alignment cannot be measured directly, it can be indirectly assessed via the phylogenetic content in the data. The phylogenetic content in our data set was evaluated by likelihood-mapping, a method that calculates ML trees for all possible quartets of sequences and counts the frequency of trees according to their quality (Strimmer and von Haeseler, 1997). A fully resolved bifurcating tree was obtained for 87.2% of all quartets, a partially resolved tree for 4.9%, while for 7.9% the tree topology could not be resolved. Thus, the phylogenetic content in the data is sufficient to infer trees that are reliable in most parts.

To further increase the reliability of the phylogenetic reconstruction, the analysis was performed with several algorithms and amino acid substitution models: NJ, minimum evolution (ME) (Kumar et al., 2001), and parsimony algorithms (Felsenstein, 1995); Bayesian probability analysis (Huelsenbeck and Ronquist, 2001) under the assumption of the DAYHOFF (Dayhoff et al., 1978) and JONES (Jones et al., 1992) substitution models; and ML analysis using quartet-puzzling (Strimmer and von Haeseler, 1996) under the assumption of the WHELAN–GOLDMAN substitution model (Whelan and Goldman, 2001).

Except of a few inconsistencies at basal branches of the non-segmented negative-strand RNA viruses, all analyses inferred the same tree topology and most of the support values ranged from 80% to 100% (Fig. 6), indicating a stable reconstruction of the phylogeny. The tree consisted of three major branches: positive-strand RNA viruses which were used to root the tree, non-segmented negative-strand RNA viruses, and segmented negative-strand RNA viruses. In the non-segmented viruses, the topology was consistent with the current taxonomic grouping into families and subfamilies, with the exception that the family *Paramyxoviridae* appeared as a paraphyletic taxon: its subfamily *Pneumovirinae* was placed in sister relationship with the family *Filoviridae* rather than with the other subfamily *Paramyxovirinae*. The relationship between the families *Borna-*, *Rhabdo-*, and *Paramyxoviridae* could not be resolved.

Within the segmented negative-strand RNA viruses, the tree topology suggests an independent evolution of two major sister lineages, the *Orthomyxoviridae* on the one hand and the *Bunyaviridae*/*Arenaviridae* on the other hand. Surprisingly, the *Arenaviridae* were placed by all methods at a distal branch of the Bunyavirus lineage in sister relationship with the nairovirus Dugbe, suggesting that *Arenaviridae* were derived from an intermediate ancestor of the *Bunyaviridae*. In conclusion, the analyses suggest that *Bunyaviridae* are a paraphyletic taxon with a close relationship between the Nairovirus genus and the *Arenaviridae* taxon, while *Orthomyxoviridae* and *Arenaviridae* are monophyletic.

Discussion

The L RNA of three Lassa virus strains originating from different regions of West Africa was sequenced, and structural and phylogenetic information was extracted from the data by an in-depth computer analysis. Database searches identified a putative RNA polymerase domain which was shown to be composed of similar secondary structures elements as found in crystal structures of viral RNA-dependent RNA polymerases. A notable exception are additional alpha helices between sequence motif A and B. A hypothetical folding model of the arenavirus polymerase domain is proposed. Phylogenies reconstructed for genes located on L and S RNA suggest that neither recombination nor reassortment took place during Lassa virus evolution. A reliable phylogeny of negative-strand RNA viruses was inferred on the basis of a structure-assisted sequence alignment.

Two features of Lassa virus L RNA made its sequence analysis difficult. First, Lassa virus L RNA was found to be highly variable in sequence as has been noted previously with the S RNA (Bowen et al., 2000). Primers designed solely on the basis of the known Lassa virus L RNA sequence often failed with the distantly related strains AV and CSF. Effective primers were eventually designed by including sequences of other arenaviruses in primer design. The three-step amplification and sequencing strategy based on these primers facilitated rapid generation of sequence data despite high variability and may be applicable in similar projects. The second problem arose in the amplification and sequencing of the stable stem-loop structure in the intergenic region as has been reported previously (Djavani et al., 1997; Salvato and Shimomaye, 1989). It could be partially overcome by applying specific reaction conditions. As observed for the Lassa virus S RNA (Auperin and McCormick, 1989; Bowen et al., 2000; Clegg et al., 1991; Günther et al., 2000, 2001), the RNA stem-loop structure of the L RNA was completely conserved, suggesting important functions of this structure in the virus life cycle.

Looking at the L protein alone, there are some clear results and some totally unresolved questions. The conserved domains identified in the L protein may roughly correspond to structural or functional domains, while regions with extremely low homology may represent spacer regions connecting these domains. The Psi-BLAST searches showed only one clearly identified and annotated region—

the RNA polymerase. Beyond this, one can only be slightly more specific. Several regions along the whole protein show a degree of conservation as found in the polymerase domain, suggesting functional or structural relevance of these regions. Both the threading and coiled-coil calculations suggest that the N terminus of the Lassa virus L protein has a high alpha-helical content and probably contains a coiled-coil domain. Another obvious query is to look for signs of expected catalytic sites outside of the RNA polymerase region. However, there is no evidence for the most obvious candidates. An exhaustive search for sequence motifs specific for viral-encoded RNA helicases of supergroup 1, 2, and 3 (Kadare and Haenni, 1997), methyltransferases (involved in RNA capping) or protein signatures as listed in the PROSITE database did not lead to any real clues. This does not mean that such sites are totally absent, but if they are present, they have evolved so as to be resistant to detection.

Within the RNA polymerase region, one can be much more specific. The similarity to other viral RNA polymerases is quite certain, but the similarity to any known structure is so remote that it is not statistically significant, even with a very rich sequence profile. Nevertheless, the obvious similarities in secondary structure mean that the structural model is probably correct in gross terms. As would be expected, the conserved sites are clustered in the centre of the proposed coordinates. The main strength of the model is the ability to locate differences to known structures. A large insertion is clear from sequence analysis alone. From the model, we can say that there is probably an extra helical segment between beta strands 5 and 6, and that it lies near to the 5′ end of the template (Fig. 4B). It may interact with nucleotide residues of the incoming template strand before they bind at the catalytic site. For example, it could be involved in recognition of the conserved RNA termini of the segmented viruses which serve as start sites for RNA transcription and replication.

A beta-loop structure in the C terminus of the HCV polymerase (strands 11 and 12) ensures that the enzyme utilises single-stranded rather than double-stranded RNA as a template and that transcription initiates only at the RNA terminus rather than internally (Hong et al., 2001; Zhong et al., 2000). The structure is essential for the function of the enzyme in the minireplicon context (Cheney et al., 2002), suggesting that recognition of and initiation at a single-strand terminus is a prerequisite for replication of the genome. An analogous element may be present in the arenavirus polymerase, although the prediction was not such reliable. Its presence would be in agreement with the current model of arenavirus replication and transcription, which is based on initiation of RNA synthesis at the RNA termini.

There is evidence that reassortment of genomic segments or recombination within a segment occurred in nature with segmented negative-strand RNA viruses such as *Bunyaviridae* and *Orthomyxoviridae* (Henderson et al., 1995; Sibold et al., 1999; Zhou et al., 1999). There is also evidence that recombination within the S RNA took place during evolution of North American arenaviruses (Charrel et al., 2002). Furthermore, closely related arenaviruses such as Lassa virus and Mopeia virus can reassort under laboratory conditions (Lukashevich, 1992). However, it is unknown whether this phenomenon occurs in nature.

The lack of evidence for reassortment or recombination in the present study is actually consistent with some opinions on Lassa virus evolution. It has been proposed that Lassa virus has evolved as it spread within West Africa from east to west (Bowen et al., 1997). However, different strains remained geographically separated thus reducing the chance of reassortment or recombination between them. Our data do not exclude that such events occurred between more closely related, cocirculating strains. But these events may not be reliably detected by phylogenetic or statistical programs.

The finding that L and Z genes show a lower level of nucleotide and amino acid sequence conservation than the structural genes NP and GPC was somewhat surprising. This phenomenon was found when comparing the genes at the level of the Lassa viruses, the Old World arenaviruses, and the whole arenavirus family. One possible explanation may be that L and Z genes have a faster rate of evolution than NP and GPC. On the other hand, the effect could be explained by the presence of several spacer regions in the L protein connecting functional domains. These regions would be more flexible in terms of sequence and may lower the overall degree of sequence conservation of the L protein. Similarly, the function of the Z protein may reside mainly in the central RING motif while other parts of the protein are not subject to major functional or structural constraints.

Reconstruction of phylogeny above the virus family level is still a challenging issue. Since RNA-dependent polymerases are encoded by a wide range of virus families and share conserved motifs in their catalytic domain, corresponding sequences appear to be particularly useful to infer the phylogeny between families. Several attempts were made with different data sets (Bruenn, 1991; Koonin, 1991; Koonin and Dolja, 1993). However, these data sets were not found to contain sufficient phylogenetic signal to infer the relationship among most of the taxa upon reevaluation using sophisticated statistics (Zanotto et al., 1996). Similarly, two phylogenies reconstructed for segmented negative-strand RNA viruses differed remarkably and the relationship between a number of taxa could not be resolved (Lukashevich et al., 1997; Marriott and Nuttall, 1996). Notably, none of these studies included structural information in the alignment.

We found that inclusion of data on secondary structure enhanced the accuracy of the amino acid sequence alignment even in regions which lack apparent sequence similarity. Using a neuronal network program for secondary structure prediction (Cuff and Barton, 2000), virtually all beta strand and helix elements of the catalytic domain could be identified and assembled into a 313-residue data set. The

data set underwent analysis by several algorithms. All of which inferred the same phylogenetic tree with generally high support values, except at basal branches of the non-segmented negative-strand viruses. Therefore, we think the phylogeny can be considered reliable. This conclusion is supported by the fact that the observed tree topology is consistent with subphylogenies previously reconstructed for Arenaviruses (Bowen et al., 1997), Bunyaviruses (Chizhikov et al., 1995; Roberts et al., 1995), and non-segmented RNA viruses (Dhillon et al., 2000). According to the obtained phylogeny, *Arenaviridae* may be considered a subgroup of the *Bunyaviridae*. Arenaviruses and Nairoviruses seem to be most closely related. The idea that these viruses share a most recent common ancestor would be consistent with the finding that the GPCs of Lassa virus and of the nairovirus Crimean–Congo hemorrhagic fever virus are cleaved by the same cellular enzyme (protease SKI-1/S1P) (Lenz et al., 2001; Vincent et al., 2003), which has so far not been implicated in glycoprotein processing of any other virus.

The results from this work suggest some future directions. Firstly, the work suggests the borders of structural domains and those parts of the sequence which would be suitable to structural, biochemical, and genetic investigation. Secondly, it highlights the areas where the arenavirus L proteins are most different to anything currently characterised. Although one could imagine RNA polymerase inhibitors, an equally profitable route to taming the virus could be through characterisation of the rest of the L protein which appears to be surprisingly unique. Thirdly, the work shows that combining structure prediction with phylogenetic analysis can be an approach to infer the phylogeny in spite of large evolutionary distances.

Materials and methods

Virus strains and RNA preparation

Propagation of Lassa virus was performed in a biosafety level 4 laboratory. Origin and isolation of Lassa virus strains AV and CSF have been described (Günther et al., 2000, 2001). Lassa virus NL was isolated from serum of a surgeon who worked in a hospital in rural Sierra Leone (Schmitz et al., 2002). For preparation of virus stocks, all strains underwent one additional passage following isolation from clinical material (i.e. second passage). The virus titer of the stocks was determined by immuno focus assay using Lassa virus NP-specific monoclonal antibody. For RNA preparation, Vero cells in four 75-cm^2 tissue culture flasks were inoculated with virus at a multiplicity of infection (MOI) of 0.01. After 4 days, supernatant was cleared by centrifugation to remove cell debris and virus was pelleted from the cleared supernatant by ultracentrifugation using a TST 28.38/17 rotor (Kontron Instruments) with 25,000 rpm at 4 °C overnight. The pellets were resuspended in 140 μl water plus 560 μl buffer AVL (Qiagen, Hilden, Germany). RNA was isolated using the viral RNA kit (Qiagen) according to the manufacturer's instructions.

RT-PCR and sequencing

Initial short-range fragments (Fig. 1, step 1) were amplified using the Superscript one-step RT-PCR system with Platinum *Taq* polymerase (Invitrogen). RT-PCR was performed in a 20-μl reaction containing 2 μl RNA, 1× buffer supplemented with 0.9 mM $MgSO_4$, 0.4 μl enzyme mix, and 0.2 μM of each plus and minus strand primer. The reaction was run in a Perkin Elmer 2400 thermocycler as follows: reverse transcription at 50 °C for 30 min; 95 °C for 5 min; touch-down amplification for 10 cycles at 95 °C for 15 s, 60 °C for 10 s with a 1 °C-decrease every cycle, and 72 °C for 1 min; amplification for 40 cycles at 95 °C for 15 s, 56 °C for 10 s, and 72 °C for 1 min; final extension at 72 °C for 5 min. When faint or multiple PCR signals appeared, the correct PCR product was gel-purified and reamplified to obtain pure and sufficient material for sequencing.

Long-range fragments (Fig. 1, step 2) were amplified similarly as described previously (Günther et al., 2000). In brief, 4 μl of purified RNA was incubated with 20 pmol of each plus and minus strand primer in an 8-μl assay at 70 °C for 15 min and quickly chilled on ice. Alternatively, only one primer was used. A 19-μl reaction premix containing 8 μl RNA-primer mix, 50 mM Tris–HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, and 500 μM dNTP was incubated at 50 °C for 2 min and subsequently 20 or 200 units Superscript II reverse transcriptase (Invitrogen) were added. The reaction was run at 50 °C for 30 min; 55 °C for 5 min; 50 °C for 20 min; 60 °C for 1 min; 50 °C for 10 min. The enzyme was inactivated at 70 °C for 15 min and RNA was removed by adding 2 units RNase H (Invitrogen) and incubating at 37 °C for 20 min. cDNA was amplified by using the Expand High Fidelity PCR System (Roche Molecular Biochemicals, Mannheim, Germany) with a hot start. The 45-μl reaction premix contained 1 μl of cDNA, 1× buffer with 1.5 mM $MgCl_2$, 200 μM dNTP, and 0.3 μM of each plus and minus strand primer. The premix was heated to 55 °C, and 5 μl enzyme mixture, containing 2.6 units *Taq* and *Pwo* polymerase in 1× buffer, was added. The PCR was run for 40 cycles with 94 °C for 1 min, 55 °C for 1.5 min, and 72 °C for 3 min with an increment of 2 min after every 10 cycles in a Robocycler (Stratagene, La Jolla, CA).

The final products for sequencing (Fig. 1, step 3) were amplified using 1 μl of long-range PCR product as a template, a nested or seminested set of primers, and 2 units *Taq* DNA polymerase (Pharmacia). The 25-μl reaction was run in a 9600 thermocycler for 40 cycles with 95 °C for 20 s, 50 °C for 30 s, and 72 °C for 1 min.

The stem-loop structure in the intergenic region was reverse transcribed using the protocol of the long-range RT-PCR as described above and amplified using 2.5 units Platinum *Taq* polymerase (Invitrogen) with standard con-

ditions. The correct (larger) fragment was purified from gel and reamplified for 25 cycles in a 50-μl reaction containing the same primers, 2.5 units Platinum *Taq* PCR_X polymerase, and 0.5× PCR_X enhancer solution for G/C-rich templates (Invitrogen). The correct fragment was again purified from gel before sequencing.

Products of the initial short-range PCR and of the final reamplification were sequenced using the PCR primers. PCR products were ethanol-precipitated and directly sequenced on both strands using an ABI 377 automated sequencer and the BigDye Terminator AmpliTaq kit (Applied Biosystems). Stem-loop fragments were sequenced using additives (5% DMSO or 10% Formamide) and higher denaturation temperatures (98 °C). Alternatively, the SequiTherm EXCEL sequencing kit (Epicentre Technologies) for LI-COR sequencer was used. Sequence coverage was 2- to 6-fold per nucleotide position. Overlapping sequences were identical. A total of 85 L RNA-specific primers was used for PCR and sequencing (see supplementary data for sequences of selected primers).

The full-length S RNA of Lassa virus NL was amplified using RNA prepared from culture supernatant and sequenced as described previously for strains AV and CSF (Günther et al., 2000, 2001).

L RNA sequences of Lassa virus AV, NL, and CSF have been sent to GenBank and assigned the accession numbers, AY179171, AY179172, AY179174, and AY179175. The S RNA sequence of strain NL has been assigned the accession number AY179173.

Sequence homology search and sequence entropy

Protein sequence database searches used the full 2220-residue sequence of Lassa strain AV as a probe against the non-redundant protein database (release date 27 April 2003) with the addition of the L genes from strains NL and CSF. Psi-BLAST (Altschul et al., 1997) was run in iterative mode in two steps. To build an initial profile, the database was searched and homologues with an expectation value less than 10^{-7} (four orders of magnitude more stringent than the default value) were included. This converged after three rounds, providing an initial profile from seven very close sequence homologues (all arenaviruses) for more than 2000 residues and a further seven close homologues for shorter fragments (all arenaviruses). This was followed by a second step using the generated profile, but with default values (accepting homologues with an expectation value $<5 \times 10^{-3}$). After 11 rounds, this converged with up to 340 homologues at some sites, including, of course, partial sequences. The final analysis of homologue data was based on those sequences with an expectation value less than 1×10^{-100}. This did select only those sequences with completely unambiguous similarity, but, more importantly, removed the large number of incomplete sequence fragments and was essential to the definition of domain boundaries.

Sequence entropy $S(k)$ for each residue position k within the Psi-BLAST alignments was calculated using a standard Shannon type information measure

$$S(k) = \sum_{t=1}^{N_t} p_t(k) \ln p_t(k)$$

where $p_t(k)$ is the probability of amino acid type t at position k and the summation runs over the N_t amino acid types. Biochemically, $N_t = 20$, but, in some calculations, a reduced alphabet was used with $N_t = 6$ types: aliphatic (A, V, L, I, M, C), aromatic (F, W, Y, H), polar (S, T, N, Q), positive (K, R), negative (D, E), and special (G, P) (Mirny and Shakhnovich, 1999).

Secondary structure prediction and modelling

Coiled-coil domains were searched for with COILS 2.1 (Lupas et al., 1991) (http://www.ch.embnet.org/software/COILS_form.html) using both the MTIDK and MTK matrices with and without weighing the hydrophobic positions. Two programs were used for threading analysis. Firstly, the L protein sequence was split into fragments of 500 residues, each overlapping by 250 residues. Each of these was given to the 3D-PSSM server (Bates et al., 2001) (http://www.sbg.bio.ic.ac.uk/~3dpssm/html/ffhome.html). Secondly, longer calculations used a local copy of the threading code WURST (Huber et al., 1999). The L protein sequence was split into fragments of 150 residues, starting every 5 residues (overlapping by 145 residues). This was repeated with fragments of length 200 residues.

Secondary structure predictions of the central L protein domain of all arenaviruses (position 1050–1490 in Lassa virus Josiah) were done with the neuronal network Jnet (Cuff and Barton, 2000) (http://www.compbio.dundee.ac.uk/~www-jpred/). In addition, secondary structure predictions were performed on the following 25 polymerase sequences (accession number): Bunyamwera virus (P20470); Dugbe virus (Q66431); Hantaan virus (P23456); Rift valley fever virus (P27316); influenza B virus (O36430); influenza C virus (P19703); influenza A virus (AAA43639); Thogoto virus (O41353); rice stripe virus (NP_620522); tomato spotted wilt virus (RRVUTW); Borna disease virus (AAA20228); bovine ephemeral fever virus (NP_065409); Ebola virus (AAD14589); Marburg virus (AAA46562); viral hemorrhagic septicemia virus (AAF04486); measles virus (AAD29091); human respiratory syncytial virus (NP_056866); human metapneumovirus (AAK62941); mumps virus (P30929); Nipah virus (NP_112028); northern cereal mosaic virus (NP_597914); rabies virus (P16289); rice yellow stunt virus (NP_620502); Sendai virus (P27566); vesicular stomatitis virus (AAA48371). The highly conserved motif C was identified in these sequences and a fragment comprising 400 amino acids upstream and 300 amino acids downstream thereof was subjected to secondary structure prediction. Homologous secondary structure ele-

ments and amino acid positions of the RNA polymerases of HCV, RHDV, and polio virus were determined by superimposing their crystal structures (PDB-ID 1C2P, 1RDR, and 1KHV, respectively) using the Swiss-PdbViewer 7.3b2 (Guex and Peitsch, 1997) (available at http://www.expasy.org/spdbv/mainpage.htm). Subsequently, the predicted elements were aligned with the elements of HCV, RHDV, and polio virus, guided by conserved motifs pre-A to E. The resulting amino acid sequence alignment was marginally refined without introducing gaps into secondary structure elements.

A three-dimensional model of the arenavirus RNA polymerase was built using 1C2P (Lesburg et al., 1999) as the template, according to the adjusted alignment with the server SWISS-MODEL (Guex and Peitsch, 1997; Peitsch, 1996) (http://www.expasy.org/swissmod/). Predicted but non-templated elements were given coordinates from ideal secondary structures. The quality of the hypothetical structure was evaluated by a set of WHAT IF checks (Hooft et al., 1996; Vriend, 1990) (at http://www.cmbi.kun.nl/gv/servers/WIWWWI/model.html). Graphics were generated with the Swiss-PdbViewer.

Phylogenetic analysis

Nucleotide and amino acid sequences of the NP, GPC, Z, and L genes were aligned by the program CLUSTALW implemented into MacVector 7.0 software (Oxford Molecular) and manually adjusted. In addition to the genes sequenced in this work, the following arenavirus sequences were included (accession number): Pichinde 3739 (K02734, AF427517); Tacaribe clone p2b-2 (M20304); Tacaribe T.RVL.II 573 (J04340); LCMV Armstrong (M20869, J04331, M27693); LCMV WE (M22138, AF004519); Lassa Josiah (J04324, U73034); Lassa AV (AF246121); Lassa CSF (AF333969). Nucleotide alignments were refined based on amino acid alignments. Identity between two aligned sequences was determined by an Omnis Studio 3.1 (Raining Data Corporation) library (kindly written by Christian Schmitz, University of Hamburg). Phylogenetic analysis of the NP, GPC, Z, and L gene alignments was performed using the PHYLIP 3.57c program package (Felsenstein, 1995) with default settings. NJ analysis was conducted using DNADIST and NEIGHBOR, and ML analysis was conducted using DNAML. Analyses were performed on a bootstrapped data set (100 replicates). Recombination was tested with the programs GENECONV 1.81 (Sawyer, 1999) and RDP 1.09 (Martin and Rybicki, 2000) (available at ftp://ftp.uct.ac.za/pub/data/geminivirus/recomb.htm#Availability).

Phylogenetic analysis of negative-strand RNA viruses was based upon the secondary structure-mediated alignment (see above). The complete alignment was trimmed by removing variable loop sequences. The final alignment of 32 sequences consisted of 313 amino acid residues of helices B, C, D, E, H, I, J, K, and L, and strands 2, 3, 4, 5, 6, 7, 8, 9, and 10 (numbering according to Fig. 4, see supplementary data for the alignment). In larger gaps, two out of three gap sites were defined as missing data; small gaps (one to three sites) were treated as a single gap site. The overall phylogenetic content of the data set was evaluated by likelihood mapping (Strimmer and von Haeseler, 1997) using TREE-PUZZLE 5.1 (available at http://www.tree-puzzle.de/) with the WHELAN–GOLDMAN substitution model (Whelan and Goldman, 2001). Phylogeny was inferred using the parsimony program PROTPARS of the PHYLIP 3.57c package (Felsenstein, 1995). Analyses were performed on a bootstrapped data set (100 replicates), and each replicate was processed 10 times in random input order using the jumble option. NJ and ME analyses were conducted with MEGA 2.1 software (Kumar et al., 2001) (available at http://www.megasoftware.net/). The analyses were run on a bootstrapped data set (1000 replicates) using the gamma distance model (gamma parameter 3) with pairwise deletion of gap sites. Bayesian inference of phylogeny was performed with the MrBayes 2.01 program (Huelsenbeck and Ronquist, 2001) (available at http://morphbank.ebc.uu.se/mrbayes/) using the substitution models JONES (Jones et al., 1992) and DAYHOFF (Dayhoff et al., 1978) with gamma distributed sites. The program was run for 120,000 generations and trees were sampled each 10 generations (2000 trees were discarded as burn-in). Analyses were run twice with identical results. A ML tree was reconstructed with TREE-PUZZLE 5.1 (Strimmer and von Haeseler, 1996) using the WHELAN–GOLDMAN substitution model (Whelan and Goldman, 2001), the gamma distance model, and 10,000 puzzling steps. Tree graphics were generated with TreeViewPPC 1.65 (available at http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Acknowledgments

The authors thank Christian Schmitz for writing a computer program. This work was supported by grants E/B31E/M0171/M5916 and E/B41G/1G309/1A403 from the Bundesamt für Wehrtechnik und Beschaffung. The Bernhard-Nocht-Institut is supported by the Bundesministerium für Gesundheit and the Freie und Hansestadt Hamburg.

References

Ago, H., Adachi, T., Yoshida, A., Yamamoto, M., Habuka, N., Yatsunami, K., Miyano, M., 1999. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. Struct. Fold. Des. 7, 1417–1426.

Akey, D.L., Malashkevich, V.N., Kim, P.S., 2001. Buried polar residues in coiled-coil interfaces. Biochemistry 40, 6352–6360.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Auperin, D.D., McCormick, J.B., 1989. Nucleotide sequence of the Lassa virus (Josiah strain) S genome RNA and amino acid sequence comparison of the N and GPC proteins to other arenaviruses. Virology 168, 421–425.

Bates, P.A., Kelley, L.A., MacCallum, R.M., Sternberg, M.J., 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins Suppl. 5, 39–46.

Borden, K.L., Campbell Dwyer, E.J., Salvato, M.S., 1998a. An arenavirus RING (zinc-binding) protein binds the oncoprotein promyelocyte leukemia protein (PML) and relocates PML nuclear bodies to the cytoplasm. J. Virol. 72, 758–766.

Borden, K.L., Campbelldwyer, E.J., Carlile, G.W., Djavani, M., Salvato, M.S., 1998b. Two RING finger proteins, the oncoprotein PML and the arenavirus Z protein, colocalize with the nuclear fraction of the ribosomal P proteins. J. Virol. 72, 3819–3826.

Bowen, M.D., Peters, C.J., Nichol, S.T., 1997. Phylogenetic analysis of the Arenaviridae: patterns of virus evolution and evidence for cospeciation between arenaviruses and their rodent hosts. Mol. Phylogenet. Evol. 8, 301–316.

Bowen, M.D., Rollin, P.E., Ksiazek, T.G., Hustad, H.L., Bausch, D.G., Demby, A.H., Bajani, M.D., Peters, C.J., Nichol, S.T., 2000. Genetic diversity among Lassa virus strains. J. Virol. 74, 6992–7004.

Bressanelli, S., Tomei, L., Roussel, A., Incitti, I., Vitale, R.L., Mathieu, M., De Francesco, R., Rey, F.A., 1999. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 96, 13034–13039.

Bruenn, J.A., 1991. Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases. Nucleic Acids Res. 19, 217–226.

Campbell Dwyer, E.J., Lai, H., MacDonald, R.C., Salvato, M.S., Borden, K.L., 2000. The lymphocytic choriomeningitis virus RING protein Z associates with eukaryotic initiation factor 4E and selectively represses translation in a RING-dependent manner. J. Virol. 74, 3293–3300.

Charrel, R.N., Feldmann, H., Fulhorst, C.F., Khelifa, R., de Chesse, R., de Lamballerie, X., 2002. Phylogeny of New World arenaviruses based on the complete coding sequences of the small genomic segment identified an evolutionary lineage produced by intrasegmental recombination. Biochem. Biophys. Res. Commun. 296, 1118–1124.

Cheney, I.W., Naim, S., Lai, V.C., Dempsey, S., Bellows, D., Walker, M.P., Shim, J.H., Horscroft, N., Hong, Z., Zhong, W., 2002. Mutations in NS5B polymerase of hepatitis C virus: impacts on in vitro enzymatic activity and viral RNA replication in the subgenomic replicon cell culture. Virology 297, 298–306.

Chizhikov, V.E., Spiropoulou, C.F., Morzunov, S.P., Monroe, M.C., Peters, C.J., Nichol, S.T., 1995. Complete genetic characterization and analysis of isolation of Sin Nombre virus. J. Virol. 69, 8132–8136.

Clegg, J.C., Wilson, S.M., Oram, J.D., 1991. Nucleotide sequence of the S RNA of Lassa virus (Nigerian strain) and comparative analysis of arenavirus gene products. Virus Res. 18, 151–164.

Cuff, J.A., Barton, G.J., 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40, 502–511.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure, vol. 5. National Biomedical Research Foundation, Washington, DC, pp. 345–352.

Dhillon, J., Cowley, J.A., Wang, Y., Walker, P.J., 2000. RNA polymerase (L) gene and genome terminal sequences of ephemeroviruses bovine ephemeral fever virus and Adelaide River virus indicate a close relationship to vesiculoviruses. Virus Res. 70, 87–95.

Djavani, M., Lukashevich, I.S., Sanchez, A., Nichol, S.T., Salvato, M.S., 1997. Completion of the Lassa fever virus sequence and identification of a RING finger open reading frame at the L RNA 5′ end. Virology 235, 414–418.

Djavani, M., Lukashevich, I.S., Salvato, M.S., 1998. Sequence comparison of the large genomic RNA segments of two strains of lymphocytic choriomeningitis virus differing in pathogenic potential for guinea pigs. Virus Genes 17, 151–155.

Felsenstein, J., 1995. PHYLIP (Phylogeny Inference Package) Version 3.57c. [Online] Department of Genetics, University of Washington, Washington, D.C. Available from: URL: http://evolution.genetics.washington.edu/phylip.html [28 January 2000, last date accessed].

Fuller-Pace, F.V., Southern, P.J., 1989. Detection of virus-specific RNA-dependent RNA polymerase activity in extracts from cells infected with lymphocytic choriomeningitis virus: in vitro synthesis of full-length viral RNA species. J. Virol. 63, 1938–1944.

Garcia, J.B., Morzunov, S.P., Levis, S., Rowe, J., Calderon, G., Enria, D., Sabattini, M., Buchmeier, M.J., Bowen, M.D., Jeor, S.C.S., 2000. Genetic diversity of the Junin virus in Argentina: geographic and temporal patterns. Virology 272, 127–136.

Guex, N., Peitsch, M.C., 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714–2723.

Günther, S., Emmerich, P., Laue, T., Kühle, O., Asper, M., Jung, A., Grewing, T., ter Meulen, J., Schmitz, H., 2000. Imported lassa fever in Germany: molecular characterization of a new lassa virus strain. Emerg. Infect. Dis. 6, 466–476.

Günther, S., Weisner, B., Roth, A., Grewing, T., Asper, M., Drosten, C., Emmerich, P., Petersen, J., Wilczek, M., Schmitz, H., 2001. Lassa fever encephalopathy: Lassa virus in cerebrospinal fluid but not in serum. J. Infect. Dis. 184, 345–349.

Hansen, J.L., Long, A.M., Schultz, S.C., 1997. Structure of the RNA-dependent RNA polymerase of poliovirus. Structure 5, 1109–1122.

Henderson, W.W., Monroe, M.C., St. Jeor, S.C., Thayer, W.P., Rowe, J.E., Peters, C.J., Nichol, S.T., 1995. Naturally occurring Sin Nombre virus genetic reassortants. Virology 214, 602–610.

Hong, Z., Cameron, C.E., Walker, M.P., Castro, C., Yao, N., Lau, J.Y., Zhong, W., 2001. A novel mechanism to ensure terminal initiation by hepatitis C virus NS5B polymerase. Virology 285, 6–11.

Hooft, R.W., Vriend, G., Sander, C., Abola, E.E., 1996. Errors in protein structures. Nature 381, 272.

Huber, T., Russell, A.J., Ayers, D., Torda, A.E., 1999. SAUSAGE: protein threading with flexible force fields. Bioinformatics 15, 1064–1065.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Iapalucci, S., Lopez, N., Rey, O., Zakin, M.M., Cohen, G.N., Franze-Fernandez, M.T., 1989a. The 5′ region of Tacaribe virus L RNA encodes a protein with a potential metal binding domain. Virology 173, 357–361.

Iapalucci, S., Lopez, R., Rey, O., Lopez, N., Franze-Fernandez, M.T., Cohen, G.N., Lucero, M., Ochoa, A., Zakin, M.M., 1989b. Tacaribe virus L gene encodes a protein of 2210 amino acid residues. Virology 170, 40–47.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282.

Kadare, G., Haenni, A.L., 1997. Virus-encoded RNA helicases. J. Virol. 71, 2583–2590.

Koonin, E.V., 1991. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J. Gen. Virol. 72, 2197–2206.

Koonin, E.V., Dolja, V.V., 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. 28, 375–430.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.

Lee, K.J., Novella, I.S., Teng, M.N., Oldstone, M.B., de La Torre, J.C., 2000. NP and L proteins of lymphocytic choriomeningitis virus (LCMV) are sufficient for efficient transcription and replication of LCMV genomic RNA analogs. J. Virol. 74, 3470–3477.

Lenz, O., ter Meulen, J., Klenk, H.D., Seidah, N.G., Garten, W., 2001. The Lassa virus glycoprotein precursor GP-C is proteolytically processed by subtilase SKI-1/S1P. Proc. Natl. Acad. Sci. U. S. A. 98, 12701–12715.

Lesburg, C.A., Cable, M.B., Ferrari, E., Hong, Z., Mannarino, A.F., Weber, P.C., 1999. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. Nat. Struct. Biol. 6, 937–943.

Lopez, N., Jacamo, R., Franze-Fernandez, M.T., 2001. Transcription and RNA replication of tacaribe virus genome and antigenome analogs require N and L proteins: Z protein is an inhibitor of these processes. J. Virol. 75, 12241–12251.

Lukashevich, I.S., 1992. Generation of reassortants between African arenaviruses. Virology 188, 600–605.

Lukashevich, I.S., Djavani, M., Shapiro, K., Sanchez, A., Ravkov, E., Nichol, S.T., Salvato, M.S., 1997. The Lassa fever virus L gene: nucleotide sequence, comparison, and precipitation of a predicted 250 kDa protein with monospecific antiserum. J. Gen. Virol. 78, 547–551.

Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting coiled coils from protein sequences. Science 252, 1162–1164.

Marriott, A.C., Nuttall, P.A., 1996. Large RNA segment of Dugbe nairovirus encodes the putative RNA polymerase. J. Gen. Virol. 77, 1775–1780.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.

Mirny, L.A., Shakhnovich, E.I., 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol. 291, 177–196.

Müller, R., Poch, O., Delarue, M., Bishop, D.H., Bouloy, M., 1994. Rift Valley fever virus L segment: correction of the sequence and possible functional role of newly identified regions conserved in RNA-dependent polymerases. J. Gen. Virol. 75, 1345–1352.

Ng, K.K., Cherney, M.M., Vazquez, A.L., Machin, A., Alonso, J.M., Parra, F., James, M.N., 2002. Crystal structures of active and inactive conformations of a caliciviral RNA-dependent RNA polymerase. J. Biol. Chem. 277, 1381–1387.

Peitsch, M.C., 1996. ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. Biochem. Soc. Trans. 24, 274–279.

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

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U. S. A. 98, 13757–13762.

Roberts, A., Rossier, C., Kolakofsky, D., Nathanson, N., Gonzalez-Scarano, F., 1995. Completion of the La Crosse virus genome sequence and genetic comparisons of the L proteins of the Bunyaviridae. Virology 206, 742–745.

Salvato, M.S., Shimomaye, E.M., 1989. The completed sequence of lymphocytic choriomeningitis virus reveals a unique RNA structure and a gene for a zinc finger protein. Virology 173, 1–10.

Salvato, M., Shimomaye, E., Oldstone, M.B., 1989. The primary structure of the lymphocytic choriomeningitis virus L gene encodes a putative RNA polymerase. Virology 169, 377–384.

Salvato, M.S., Schweighofer, K.J., Burns, J., Shimomaye, E.M., 1992. Biochemical and immunological evidence that the 11 kDa zinc-binding protein of lymphocytic choriomeningitis virus is a structural component of the virus. Virus Res. 22, 185–198.

Sawyer, S.A., 1999. GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

Schmitz, H., Köhler, B., Laue, T., Drosten, C., Veldkamp, P.J., Günther, S., Emmerich, P., Geisen, H.P., Fleischer, K., Beersma, M.F., Hoerauf, A., 2002. Monitoring of clinical and laboratory data in two cases of imported Lassa fever. Microbes Infect. 4, 43–50.

Sibold, C., Meisel, H., Krüger, D.H., Labuda, M., Lysy, J., Kozuch, O., Pejcoch, M., Vaheri, A., Plyusnin, A., 1999. Recombination in Tula hantavirus evolution: analysis of genetic lineages from Slovakia. J. Virol. 73, 667–675.

Singh, M.K., Fuller-Pace, F.V., Buchmeier, M.J., Southern, P.J., 1987. Analysis of the genomic L RNA segment from lymphocytic choriomeningitis virus. Virology 161, 448–456.

Strimmer, K., von Haeseler, A., 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13, 964–969.

Strimmer, K., von Haeseler, A., 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc. Natl. Acad. Sci. U. S. A. 94, 6815–6819.

Vincent, M.J., Sanchez, A.J., Erickson, B.R., Basak, A., Chretien, M., Seidah, N.G., Nichol, S.T., 2003. Crimean–Congo hemorrhagic fever virus glycoprotein proteolytic processing by subtilase SKI-1. J. Virol. 77, 8640–8649.

Vriend, G., 1990. WHAT IF: a molecular modeling and drug design program. J. Mol. Graph. 8 (52–6), 29.

Weaver, S.C., Salas, R.A., de Manzione, N., Fulhorst, C.F., Duno, G., Utrera, A., Mills, J.N., Ksiazek, T.G., Tovar, D., Tesh, R.B., 2000. Guanarito virus (Arenaviridae) isolates from endemic and outlying localities in Venezuela: sequence comparisons among and within strains isolated from Venezuelan hemorrhagic fever patients and rodents. Virology 266, 189–195.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Zanotto, P.M., Gibbs, M.J., Gould, E.A., Holmes, E.C., 1996. A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J. Virol. 70, 6083–6096.

Zhong, W., Ferrari, E., Lesburg, C.A., Maag, D., Ghosh, S.K., Cameron, C.E., Lau, J.Y., Hong, Z., 2000. Template/primer requirements and single nucleotide incorporation by hepatitis C virus nonstructural protein 5B polymerase. J. Virol. 74, 9134–9143.

Zhou, N.N., Senne, D.A., Landgraf, J.S., Swenson, S.L., Erickson, G., Rossow, K., Liu, L., Yoon, K., Krauss, S., Webster, R.G., 1999. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. J. Virol. 73, 8851–8856.