Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 24–28, January 1985
Biochemistry

Sequences of the *S1* genes of the three serotypes of reovirus

(cloning/homology/evolution/genetic relatedness/reading frames)

L. WILLIAM CASHDOLLAR*, RICHARD A. CHMELO, JON R. WIENER, AND WOLFGANG K. JOKLIK†

Department of Microbiology and Immunology, Duke University Medical Center, Durham, NC 27710

Contributed by Wolfgang K. Joklik, August 20, 1984

ABSTRACT **The *S1* genes of the three serotypes of reovirus have been cloned and sequenced. The *S1* genes encode protein σ1, the protein against which serotype-specific neutralizing antibodies are directed; it is also the reovirus hemagglutinin and cell-attachment protein and is a major determinant of host range/tissue specificity and of the nature of the interaction of reovirus with cells of the immune system. The *S1* genes of serotypes 1, 2, and 3 are 1458, 1442, and 1416 nucleotides long, respectively. They possess untranslated regions 13, 13, and 12 nucleotides long at their 5′ termini and 188, 229, and 36 nucleotides long at their 3′ termini. They possess two open reading frames. The first starts with a "weak" initiation codon and extends for 418, 399, and 455 codons, respectively; this is the size expected for the σ1 proteins. The other reading frame starts at a "strong" initiation codon about 70 residues downstream from the 5′ terminus but extends for only about 120 codons, being terminated by 3 in-phase termination codons in all three genes. The proteins encoded by these short open reading frames are quite basic. The serotype 1 and 2 *S1* genes are much more closely related to each other (28% homology) than to the serotype 3 *S1* gene (5% and 9% homology, respectively). These figures are based on direct homology calculations, adjusted for 25% random coincidence. Serologic evidence and hydrophobicity profiles agree that the σ1 proteins of serotypes 1 and 2 are much more closely related to each other (about 40% homology) than to that of serotype 3 (only about 20% homology). The fact that the serotype 1 and 2 *S1* genes are much more closely related to each other than to the serotype 3 *S1* gene is remarkable since for all other nine reovirus genes the serotype 1 and 3 genes are much more closely related to each other than to the serotype 2 gene. Mechanisms that may effect this remarkable evolutionary pattern are discussed.**

There are three serotypes of reovirus. As judged by the ability of their double-stranded RNAs to hybridize with each other under standardized conditions, serotypes 1 and 3 are closely related (about 70%), whereas serotype 2 is related to serotypes 1 and 3 only to the extent of about 10% (1). The individual genes of the three serotypes vary in their relatedness from those that are most closely related, the 3 *L* genes, to the gene that has diverged to the greatest extent during evolution—namely, the *S1* gene, which encodes protein σ1 (1). Protein σ1 is not only the most type-specific of all reovirus proteins but is also functionally of great importance. It is the protein against which the major neutralizing antibodies are formed (2); it is the hemagglutinin (3); it is the viral cell-attachment protein (4); it is the major determinant of the nature of the interaction of reoviruses with cells, including cells of the immune system (5–7); and it also appears to be the reovirus protein that switches off host cell DNA synthesis (8).

Remarkably, protein σ1 is present in reovirus particles to the extent of only 24 molecules that are located pairwise at 12 positions on the reovirus surface where the core projections or spikes protrude through the outer capsid shell (4). Not only is protein σ1 only a minor reovirus particle component but it is also synthesized in infected cells in only small amounts; its mRNA is one of the most inefficiently translated reovirus mRNAs [>50 times less efficiently than the most efficiently translated reovirus mRNA—namely, s4 RNA (unpublished results)].

We have devised a technique for cloning reovirus genes into pBR322 (9). We present here the sequences of the *S1* genes of all three reovirus serotypes, together with the deduced sequences of the σ1 proteins that they encode.

MATERIALS AND METHODS

The method used for growing the Lang strain of reovirus serotype 1, the D5/Jones strain of serotype 2, and the Dearing strain of serotype 3 was that described by Smith *et al.* (10). The procedures used for extracting reovirus genomic RNA, transcribing it into full-length cDNA, and cloning the double-stranded cDNA into pBR322 have been described (9, 11). Total genomic RNA of each serotype was used as the template for preparing cDNA. Before cloning, the populations of single-stranded cDNA molecules were enriched for full-length *S1* gene transcripts by electrophoresis in alkaline agarose gels (11). The cloned *S1* genes were sequenced by subcloning various restriction fragments into M13 mp8 and mp9 vectors (New England Biolabs) (12) and using the chain-terminator method of Sanger *et al.* (13). All regions of all 3 genes were sequenced in both directions.

RESULTS

Sequences of the 3 *S1* Genes. The sequences of the *S1* genes of reovirus serotypes 1–3 are presented in Fig. 1. The 3 genes are 1458, 1442, and 1416 base pairs (bp) long, respectively. The first initiation codons start at positions 14, 14, and 13, respectively, and are followed by open reading frames 1254, 1197, and 1365 bp long, capable of encoding proteins 418, 399, and 455 amino acids long, respectively (Table 1). These are the sizes expected of the σ1 proteins, based on electrophoretic migration rates in $NaDodSO_4$/polyacrylamide gels.

There is another initiation codon, in a different reading frame, that starts at residues 75, 66, and 71, respectively, in all 3 *S1* genes (Fig. 1). These are "strong" initiation codons (14), with purines in positions −3 and +4; by contrast, the upstream codons are all "weak" (Table 1). However, the σ1 proteins are translated from the first set of initiation codons and not from the second, because the latter are followed by reading frames that are only 119, 125, and 120 codons long, respectively. Interestingly, these short reading frames are all terminated by 3 in-phase termination codons; they are clearly highly conserved.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: bp, base pair(s).

*Present address: Department of Microbiology, Medical College of Wisconsin, Milwaukee, WI 53226.
†To whom reprint requests should be addressed.

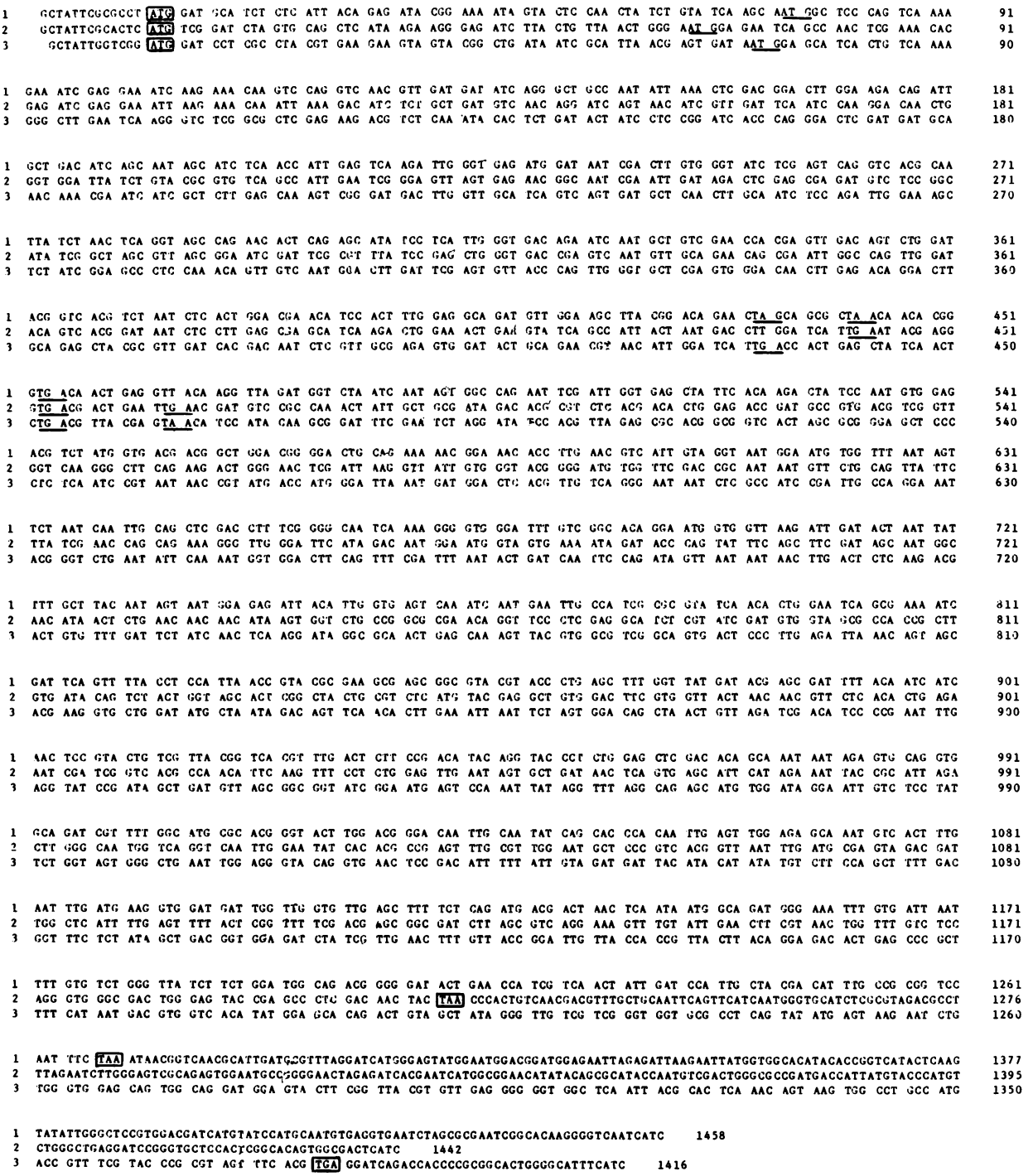

FIG. 1. Sequences of the serotype 1–3 *S1* genes of reovirus. The sequences are those of the plus (mRNA) strands.

Relatedness of the 3 *S1* Genes. The relatedness of the 3 *S1* genes was determined by assessment of base/base identities after locating possible deletions and insertions in each gene and matching cognate sequences. The results are presented in Table 2. The *S1* genes of serotypes 1 and 2 are much more closely related to each other than to that of serotype 3; the absolute homologies of the *S1* gene of serotype 3 with those of serotypes 1 and 2 exceed the 25% random matching level

Table 1. Organization of the *S1* genes of the three reovirus serotypes

Serotype	Length of 5′-untranslated sequences, bp	Sequences around first initiation codon	Length of long reading frame, bp	Position of second initiation codon	Sequences around second initiation codon	Length of short reading frame, bp	Length of 3′-untranslated sequences, bp
1	13	*C*CT ATG *G*	1254	75*	*G*CA ATG *G*	357	188
2	13	*C*TC ATG *T*	1197	66	*G*GA ATG *G*	375	229
3	12	*C*GG ATG *G*	1365	71	*A*TA ATG *G*	360	36

*This is the third initiation codon. The second initiation codon is followed by a termination codon 10 codons downstream.

Table 2. Relatedness of the 3 *S1* genes of reovirus

Serotypes	% homology		
	Long reading frame	Conserved portion of 3'-terminal untranslated region	σ1 proteins
1:2	28	46	36
1:3	5	21	16
2:3	9	27	14

All values are corrected for random coincidences (25% for nucleic acid and 5% for protein sequences, respectively).

by only 5% and 9%, respectively. The homologies of the long and short open reading frames are similar. Interestingly, there is a rather highly conserved 96-bp sequence near the 3' terminus of the plus strand of all 3 genes (Fig. 2 and Table 2); for this sequence the direct homology values, uncorrected for random coincidences, range from almost 50% to >70%. The significance of this high degree of relatedness is not clear, especially since only in the serotype 3 gene is this sequence in the coding region.

Amino Acid Sequences of the Three σ1 Proteins. The amino acid sequences of the three σ1 proteins are presented in Fig. 3. There is significant homology between all these proteins provided that account is taken of several deletions/insertions. The sequences of the serotype 1 and 2 proteins are remarkably congruent; there is only one deletion or insertion of 7 amino acids starting at position 150 and two minor deletions/insertions near the COOH terminus. The sequence of the serotype 3 protein is only slightly less congruent; relative to the sequence of the serotype 1 protein, it contains an insertion of 4 amino acids at positions 33–36, a deletion of 14 amino acids starting at residue 153, and a deletion of 7 amino acids starting at position 316.

When cognate sequences are brought into apposition in this manner, the serotype 1 and 2 σ1 proteins are 41% homologous with each other and 21% and 19%, respectively, with the σ1 protein of serotype 3. Forty-three amino acids (slightly more than 10%) are shared by the σ1 proteins of all three serotypes.

The amino acid compositions of the three σ1 proteins are unremarkable. They contain about the same number of hydrophobic and uncharged polar amino acids and of basic and acidic amino acids. The serotype 1 σ1 protein lacks cysteine, and the others contain only one molecule of cysteine. Thus, the pairs of σ1 protein molecules that exist in reovirus particles are unlikely to be -SS-linked dimers.

The hydrophobicity profiles of the three σ1 proteins are shown in Fig. 4. The serotype 1 and 2 σ1 protein profiles are seen to be rather similar, but that for the serotype 3 protein is quite different, except in the terminal regions. It is conceivable that the domain for the conserved σ1 function—namely, ability to bind to cells—is located in one or other of the terminal regions.

Molecular Properties of the Proteins Encoded by the Short Reading Frames. The short reading frames of the *S1* genes of serotypes 1, 2, and 3 are capable of encoding proteins that are 119, 125, and 120 amino acids long, respectively. These proteins are basic: they contain 29–35 lysine and arginine residues but only 10–13 aspartic and glutamic acid residues. They contain 46–49 nonpolar and 39–44 uncharged polar amino acids. The proteins of serotypes 1 and 2, 1 and 3, and 2 and 3 share 35%, 17%, and 27% homology, respectively.

DISCUSSION

The *S1* gene is the reovirus gene that has diverged most markedly during evolution (1). The serotype 1 and 2 *S1* genes are still related, after taking account of sequence deletions and insertions, to the extent of 53%, but the serotype 3 *S1* gene is related to the other two to the extent of only 30% and 34%, respectively. After adjusting for a background of 25% for random matching, these values reduce to absolute homology values of 28%, 5%, and 9%, respectively.

These gene homology relationships are reflected in those of the three σ1 proteins. The serotype 1 and 2 σ1 proteins share 41% of their amino acid residues. Many of the amino acid substitutions are conservative and the hydrophobicity profiles of the two proteins are very similar. The serotype 3 σ1 protein, on the other hand, is very different. It shares no more than 20% of its amino acids with either the serotype 1 or the serotype 2 σ1 protein, and its hydrophobicity profile is significantly different from that of the other two.

The reading frames from which the σ1 proteins are translated are clearly the long open reading frames that start at residue 13 or 14. The fact that the initiation codons at the beginning of these reading frames are weak is in accord with the fact that the S1 mRNAs are translated poorly and that only small amounts of the σ1 proteins are formed in infected cells (16). By contrast, the initiation codons that open up the short open reading frame are strong, and ribosome binding protection (17) and f-Met dipeptide formation experiments (18) indicate that they are functional. It should be possible to identify these proteins in infected cells, possibly by the use of antisera directed against certain regions of these putative proteins.

A curious feature of the σ1 protein of serotype 3 is its length; it is the largest of the three σ1 proteins, but in every $NaDodSO_4$/polyacrylamide system that has been used—in particular, in the phosphate-based system of Zweerink *et al.* (19) and in the Tris/glycine systems of Laemmli (20) and Maizel (21)—it migrates more rapidly than the σ1 proteins of serotypes 1 and 2. In fact, in the Maizel system the serotype 3 σ1 protein migrates faster than the serotype 3 σ2 protein, the molecular weight of which is known to be about 38,000 (9); in the other two systems its molecular weight, compared with that of a variety of marker proteins, is about 42,000 (10, 19). However, its actual molecular weight is almost 48,000. It has been suggested that the σ1 protein may be associated with the plasma membrane (22); but the σ1 protein in mature reovirus particles does not appear to be a cleavage product of a precursor because the size of the free form of protein σ1 in infected cells is the same as that of the form that exists in mature virus particles (19). Thus, the reason for the size discrepancy remains to be discovered.

The various monoclonal antibodies that have been raised against the σ1 protein of reovirus serotype 3 differ in their relative ability to neutralize infectivity on the one hand and inhibit hemagglutination (that is, cell binding) on the other (23, 24). This suggests that the σ1 proteins comprise at least two domains—one containing the dominant epitope and the

```
Serotype
   1    AATGGACGGATGGAGAATTAGAGATTAAGAATTATGGTGGCACATACACCGGTCATACTCAAGTATATTGGGCTCCGTGGACGATCATGTATCCATC    1411
   2    .G....AT.CC..G...C.......C.C....C....C..A.....T..A.CG.....CA.T..CG.C.....G...AT...C..T.....C.....    1394
   3    .G...CA........T.C.TCG.T.ACGTGT.G.G..G..TGGC.CA.TTACG..CT.AA.CAGTA.G...C..G.CAT...CG.TTC...C..GC.    1367
```

FIG. 2. Sequences of conserved regions near the 3' termini of the plus strands of the 3 *S1* genes. Identity is indicated by periods. The sequences are 96 nucleotides long.

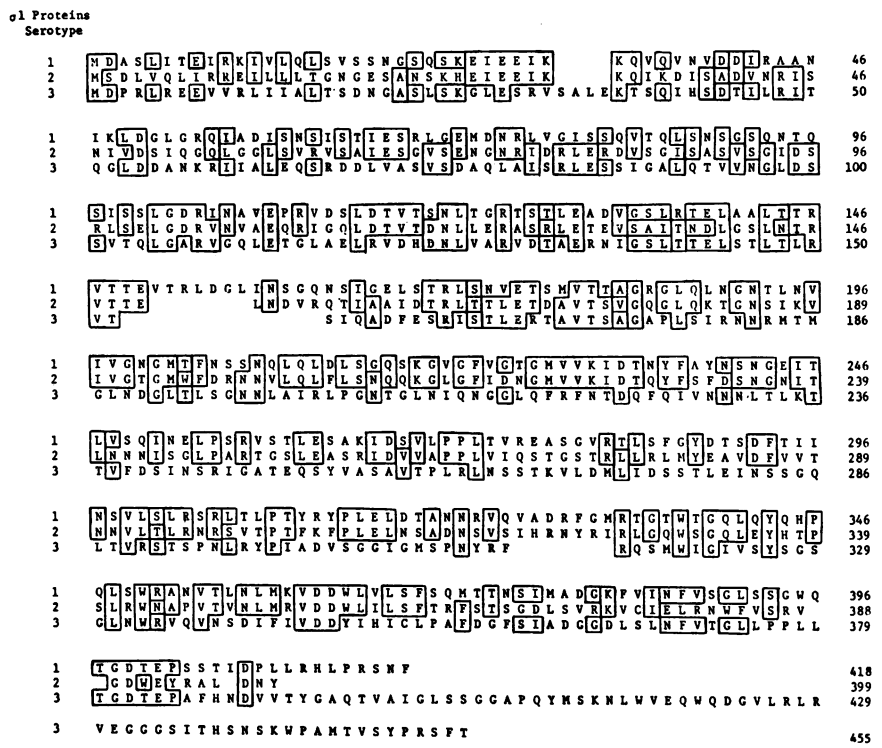

FIG. 3. Sequences of the σ1 proteins of reovirus serotypes 1–3.

other containing the region with affinity for cell receptors. Further, the amino acid residues that are shared between the three σ1 proteins tend to be clustered in the terminal regions (see Fig. 3); this is indicative of conservation of function and also suggests the existence of domains. Availability of the sequences of the three σ1 proteins opens up several approaches for detailed studies concerning their functions.

A remarkable aspect of the 3 *S1* genes is the nature of their genetic relatedness pattern. The total genomes of serotypes 1 and 3 are far more closely related to each other than to that of serotype 2 (1). All individual genes show the same pattern: all individual genes of serotypes 1 and 3 are much more closely related to each other than to those of serotype 2. The only exceptions are the 3 *S1* genes for which the sequence

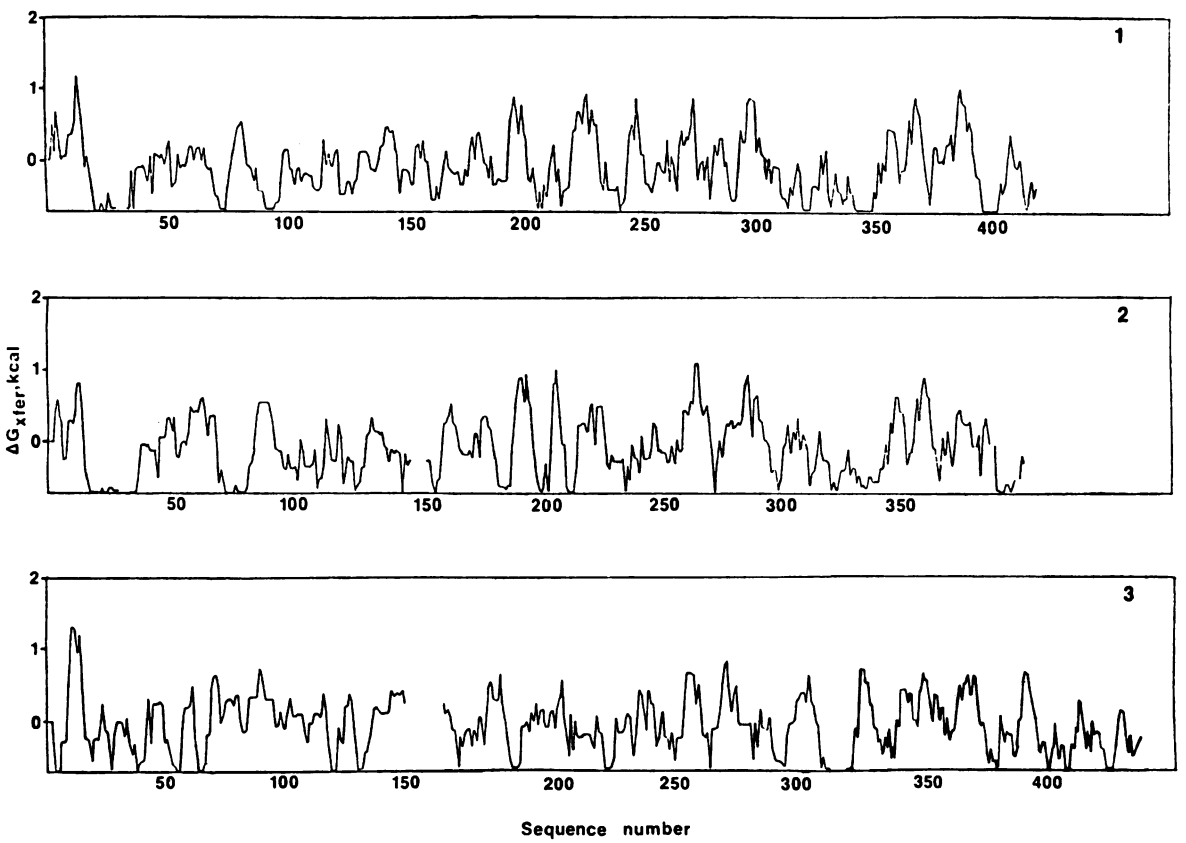

FIG. 4. Hydrophobicity profiles (15) of the σ1 proteins of reovirus serotypes 1–3.

analysis presented here demonstrates that the genes of serotypes 1 and 2 are much more closely related to each other than to that of serotype 3. This agrees with the antigenic evidence that antibodies against serotype 3 protein σ1 are absolutely specific, whereas those against serotype 1 and 2 σ1 proteins are partly cross-reactive (25). Thus, 9 of the 10 reovirus genes show a serotype 1:3 relatedness pattern, whereas the *S1* gene exhibits serotype 1:2 relatedness. Perhaps the ready occurrence of gene reassortment is responsible for this pattern of evolution, for one possible explanation for it is that at some time during evolution the *S1* genes of serotypes 2 and 3 became associated with each other's (that is, the heterologous) gene pools. It is conceivable that the evolution of an ancestral reovirus did indeed proceed along three independent major pathways, yielding three independent gene pools (1), with the *S1* gene diverging most rapidly because of selective immunologic pressures on the one hand and tolerance of extensive structural diversification on the other (because protein σ1 does not appear to be a true capsomer component, in contrast to the other two reovirus outer shell components, μ1C and σ3). Reassortants would be formed readily but be selected against because heterologous capsids would be less stable than homologous ones. At some relatively recent stage of evolution, exchange of serotype 2 and 3 *S1* genes might then have been favored by some selective mechanism like improved stability of heterologous σ1–capsid combinations caused by a structural modification, or the characteristics of the available immunologic pressures, perhaps modified by the acquisition of new host species.

We gratefully acknowledge the expert technical assistance of Patricia McClellan-Green and Bruce Macdonald. This work was supported by Research Grant 1 RO1 08909 and Program Project Grant 2 PO1 CA 30246 from the National Institutes of Health.

1. Gaillard, R. K. & Joklik, W. K. (1982) *Virology* **123,** 152–164.
2. Weiner, H. L. & Fields, B. N. (1977) *J. Exp. Med.* **146,** 1305–1310.
3. Weiner, H. L., Ramig, R. F., Mustoe, T. A. & Fields, B. N. (1978) *Virology* **86,** 581–584.
4. Lee, P. W. K., Hayes, E. C. & Joklik, W. K. (1981) *Virology* **108,** 156–163.
5. Weiner, H. L., Drayna, D., Averill, D. R., Jr., & Fields, B. N. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5744–5748.
6. Fontana, A. & Weiner, H. L. (1980) *J. Immunol.* **125,** 2660–2664.
7. Weiner, H. L., Greene, M. I. & Fields, B. N. (1980) *J. Immunol.* **125,** 278–282.
8. Sharpe, A. H. & Fields, B. N. (1981) *J. Virol.* **38,** 389–392.
9. Cashdollar, L. W., Esparza, J., Hudson, G. R., Chmelo, R., Lee, P. W. K. & Joklik, W. K. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 7644–7648.
10. Smith, R. E., Zweerink, H. J. & Joklik, W. K. (1969) *Virology* **39,** 791–810.
11. Cashdollar, L. W., Chmelo, R., Esparza, J., Hudson, G. R. & Joklik, W. K. (1984) *Virology* **133,** 191–196.
12. Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9,** 309–321.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
14. Kozak, M. (1981) *Nucleic Acids Res.* **9,** 5233–5252.
15. Rose, G. D. & Roy, S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 4643–4647.
16. Zweerink, H. J. & Joklik, W. K. (1970) *Virology* **41,** 501–518.
17. Kozak, M. (1982) *J. Mol. Biol.* **156,** 807–820.
18. Cenatiempo, Y., Twardowski, T., Shoeman, R., Ernst, H., Brot, N., Weissbach, H. & Shatkin, A. J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1084–1088.
19. Zweerink, H. J., McDowell, M. J. & Joklik, W. K. (1971) *Virology* **45,** 716–723.
20. Laemmli, U. K. (1970) *Nature (London)* **227,** 680–685.
21. Maizel, J. V. (1971) *Methods Virol.* **5,** 177–246.
22. Kauffman, R. S., Lee, S. & Finberg, R. (1983) *Virology* **131,** 265–273.
23. Hayes, E. C., Lee, P. W. K., Miller, S. E. & Joklik, W. K. (1981) *Virology* **108,** 147–155.
24. Spriggs, D. R., Kaye, K. & Fields, B. N. (1983) *Virology* **127,** 220–224.
25. Gaillard, R. K. & Joklik, W. K. (1980) *Virology* **107,** 533–536.