Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus

John Armstrong, Heiner Niemann*, Sjef Smeekens†‡, Peter Rottier† & Graham Warren

European Molecular Biology Laboratory, Postfach 10.2209,
6900 Heidelberg, FRG
* Institut für Virologie, Fachbereich Humanmedizin der Justus-Liebig Universität, Giessen, FRG
† Institute of Virology, Veterinary Faculty, State University of Utrecht, 3509 TD Utrecht, The Netherlands

In the eukaryotic cell, both secreted and plasma membrane proteins are synthesized at the endoplasmic reticulum, then transported, via the Golgi complex, to the cell surface[1–4]. Each of the compartments of this transport pathway carries out particular metabolic functions[5–8], and therefore presumably contains a distinct complement of membrane proteins. Thus, mechanisms must exist for localizing such proteins to their respective destinations. However, a major obstacle to the study of such mechanisms is that the isolation and detailed analysis of such internal membrane proteins pose formidable technical problems. We have therefore used the E1 glycoprotein from coronavirus MHV-A59 as a viral model for this class of protein. Here we present the primary structure of the protein, determined by analysis of cDNA clones prepared from viral mRNA. In combination with a previous study of its assembly into the endoplasmic reticulum membrane[9], the sequence reveals several unusual features of the protein which may be related to its intracellular localization.

The coronaviruses are a diverse class of enveloped RNA viruses of considerable medical and agricultural significance; they also provide a model for the study of persistent viral infections (see ref. 10 for review). In contrast to many enveloped viruses, the coronavirus mouse hepatitis virus (MHV) A59 buds inside the cell, into the lumen of the endoplasmic reticulum[11–14]. The assembled virion then appears to travel, via the Golgi complex, to the cell surface. Of the two viral membrane proteins, the smaller one, E1, is necessary for formation of the envelope, and is restricted to internal cell membranes; apparently it only reaches the cell surface as part of the budded virion[12,13]. Thus, the E1 glycoprotein is potentially a convenient model for studying those features of a membrane protein that determine its arrest at a particular destination on the membrane transport pathway.

The mRNAs of MHV-A59 form a 'nested set': the seven RNAs share the 3′ region of the positive-stranded genome, but extend to different lengths towards the 5′ end[15–18]. From each RNA, only the 5′ gene is translated[19,20]. In addition, a non-coding 'leader' sequence of approximately 70 bases, from the 5′ end of the genome, is common to the mRNAs[18,21,22]. The E1 gene is second from the 3′ end and is therefore translated from the second smallest mRNA, RNA 6 (refs 19, 20). The sequence of the 3′-terminal gene, encoding the viral nucleocapsid protein, has been determined previously[23,24].

Copy DNA clones spanning the E1 gene were prepared by two methods[23–25] and sequenced in the vectors M13mp8 (ref. 26) or pEMBL8 (ref. 27) by the chain-termination method[28] (data available on request). A sequence of 780 nucleotides (Fig. 1), containing a single long open reading frame, precedes the coding region for the viral nucleocapsid protein. A leader of 76 nucleotides, almost identical to the leader of the smallest mRNA, RNA 7 (ref. 22), lies in front of the first potential initiator codon. Thus, the sequence in Fig. 1 represents the 5′ end of RNA 6, encoding the E1 protein and starting at or near the extreme 5′-terminal nucleotide.

```
CCTATAAGAGTGATTGGCGTCCGTACGTACCCTCTCAACTCTAAAACTCTTGTAGTTTAA
        10                  30                  50

                MetSerSerThrThrGlnAlaProGluProValTyrGlnTrpTh
ATCTAATCCAAACATTATGAGTAGTACTACTCAGGCCCCAGAGCCCGTCTATCAATGGAC
        70                  90                 110

rAlaAspGluAlaValGlnPheLeuLysGluTrpAsnPheSerLeuGlyIleIleLeuLe
GGCCGACGAGGCAGTTCAATTCCTTAAGGAATGGAACTTCTCGTTGGGCATTATACTACT
       130                 150                 170

uPheIleThrIleIleLeuGlnPheGlyTyrThrSerArgSerMetPheIleTyrValVa
CTTTATTACTATCATACTACAGTTCGGTTACACGAGCCGTAGCATGTTTATTTATGTTGT
       190                 210                 230

lLysMetIleIleLeuTrpLeuMetTrpProLeuThrIleValLeuCysIlePheAsnCy
GAAAATGATAATCTTGTGGTTAATGTGGCCACTGACTATTGTTTTGTGTATTTTCAATTG
       250                 270                 290

sValTyrAlaLeuAsnAsnValTyrLeuGlyPheSerIleValPheThrIleValSerIl
CGTGTATGCGCTAAATAATGTGTATCTTGGATTTTCTATAGTGTTTACTATAGTGTCCAT
       310                 330                 350

eValIleTrpIleMetTyrPheValAsnSerIleArgLeuPheIleArgThrGlySerTr
TGTAATCTGGATTATGTATTTTGTTAATAGCATAAGGTTGTTTATCAGGACTGGTAGCTG
       370                 390                 410

pTrpSerPheAsnProGluThrAsnAsnLeuMetCysIleAspMetLysGlyThrValTy
GTGGAGCTTCAACCCCGAAACAAACAACCTTATGTGTATAGATATGAAAGGTACCGTGTA
       430                 450                 470

rValArgProIleIleGluAspTyrHisThrLeuThrAlaThrIleIleArgGlyHisLe
TGTTAGACCCATTATTGAGGATTACCATACACTAACAGCCACTATTATTCGTGGCCACCT
       490                 510                 530

uTyrMetGlnGlyValLysLeuGlyThrGlyPheSerLeuSerAspLeuProAlaTyrVa
CTACATGCAAGGTGTTAAGCTAGGCACCGGTTTCTCTTTGTCTGACTTGCCCGCTTATGT
       550                 570                 590

lThrValAlaLysValSerHisLeuCysThrTyrLysArgAlaPheLeuAspLysValAs
TACAGTTGCTAAGGTGTCACACCTTTGCACTTATAAGCGCGCATTCTTAGACAAGGTAGA
       610                 630                 650

pGlyValSerGlyPheAlaValTyrValLysSerLysValGlyAsnTyrArgLeuProSe
CGGTGTTAGCGGTTTTGCTGTTTATGTGAAGTCCAAGGTCGGAAATTACCGACTGCCCTC
       670                 690                 710

rAsnLysProSerGlyAlaAspThrAlaLeuLeuArgIle*
AAACAAACCGAGTGGCGCGGACACCGCATTGTTGAGAATCTAATCTAAACTTTAAGGATG
       730                 750                 770
```

Fig. 1 Sequence of the E1 cDNA and protein extending to the initiator codon of the adjacent nucleocapsid gene[23]. Proposed membrane-spanning regions are overlined.

Two versions were found, in two different clones, for the sequence immediately upstream from the E1 initiator codon. The shorter one is shown in Fig. 1; in the second clone, an additional copy of the pentanucleotide ATCTA was found between nucleotides 65 and 66, making the sequence similar to that of the region adjacent to the nucleocapsid gene of another strain of MHV[29]. This difference could represent a mutation; alternatively, it may reflect heterogeneity in the normal mRNA population. Indirect support for the latter possibility comes from the observation that a RNase-T$_1$ oligonucleotide from this region of RNA 6, corresponding to the shorter sequence, was recovered in markedly lower yield than those from the rest of the molecule[30]. This site represents the point of fusion between the 5′ leader sequence and the coding portion of the RNA. The fusion is thought to occur by 'jumping' of the viral RNA polymerase to particular sites on its genome-length, negative-stranded template; the resumption of transcription then produces each of the subgenomic mRNAs[22,31,32]. Thus, it seems possible that the polymerase may jump to more than one point on the template for each mRNA, generating variable numbers of the repeated pentanucleotide AUCUA in the resulting transcript.

Figure 1 shows the amino acid sequence encoded by the E1 gene. The predicted molecular weight of the protein is 26,000, slightly higher than that observed by gel electrophoresis[19,33] but consistent with the unusual electrophoretic behaviour of this[33], and other, hydrophobic proteins. Several features of the protein, when assembled into membranes in the virus[33], or *in vitro*[9], are reflected in the sequence. First, in contrast to the majority of membrane proteins, E1 is known to lack a cleaved 'signal peptide'[9]: the N-terminal region of the sequence contains no good candidate for a cleavage site[34]. Second, the N-terminal region bears *O*-linked sugars[35,36], which, uniquely among viral proteins so far studied, are the only known post-translational modification to E1. Assuming that the terminal Met is removed[37], the N-terminal sequence is Ser-Ser-Thr-Thr, which is identical to the *O*-glycosylated amino terminus of M-type glycophorin A (ref. 38). The *O*-linked sugars of E1 are them-

‡ Present address: Molecular Cell Biology Group, State University of Utrecht, 3509 CH Utrecht, The Netherlands.

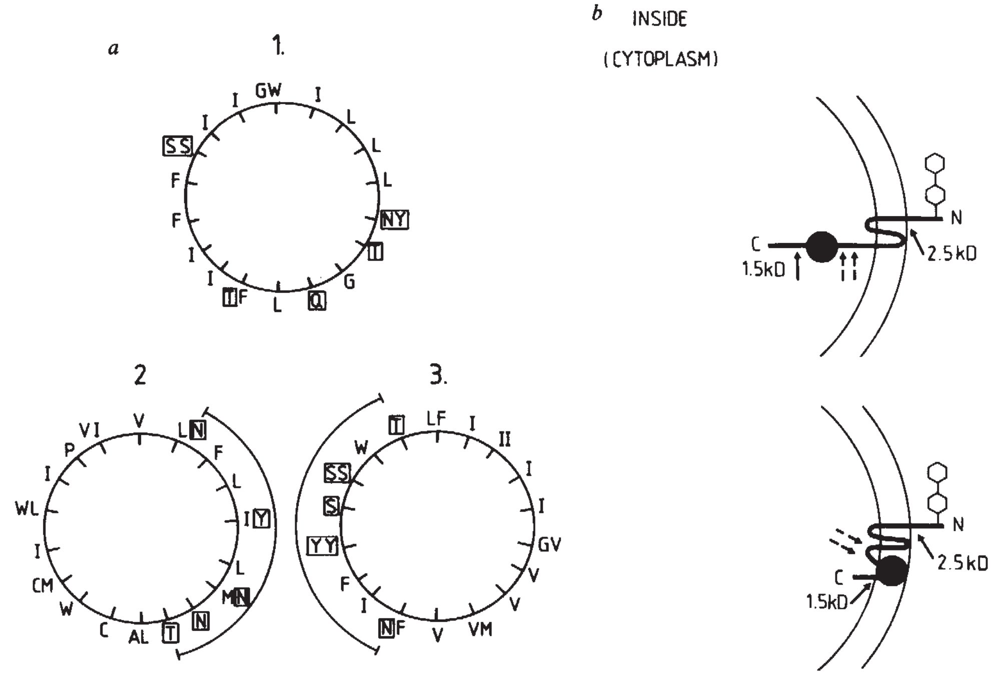

Fig. 2 *a*, Distribution of polar side chains in the hydrophobic regions of the E1 sequence. Residues 26–47 (1), 57–81 (2) and 82–106 (3) are plotted as α-helices and viewed end-on. Polar side chains are boxed: proposed hydrophilic faces of helices 2 and 3 are indicated. *b*, Possible topologies of the E1 protein across the membrane. Arrows indicate sites accessible to protease; broken arrows represent inefficient proteolysis[9].

selves identical to those found in glycophorin[39]. Third, most of the protein is resistant to proteolysis when assembled in the membrane. Only 2.5 kilodaltons of polypeptide from the N-terminus are cleavable on the luminal side of the membrane (or outside the virion) and 1.5 kilodaltons from the C-terminus from the cytoplasmic (or intra-virion) side[9], suggesting that the protein is largely buried in the membrane. In the sequence, a run of 22 uncharged residues from positions 26 to 47 represents a potential membrane-spanning region; residues 1–25 correspond to the portion removable by protease. A further sequence of uncharged residues, positions 57–106, is sufficiently long to cross the membrane twice more. If this region is divided in two, and each half plotted as an α-helical 'wheel', all the polar side chains of both sections cluster within 140°. Thus, a plausible conformation for this region is two hairpinned helices in the membrane, with adjacent polar faces (Fig. 2*a*). There are no other long hydrophobic sequences, implying that the region from residues 107 to ~190 is either folded in the membrane to neutralize charges, or, more likely, is adjacent to the membrane but resistant to proteolysis. These features are summarized in Fig. 2*b*.

Which, if any, of these various features might be responsible for the protein's intracellular localization? We do not know, for example, whether the protein has an active 'signal' causing its arrest on the transport pathway, or, alternatively, if it lacks a signal for onward transport; nor do we know whether a sorting process might operate on one or the other side of the membrane. The availability of a cDNA clone for the protein presents the opportunity to investigate these questions by allowing expression of the cloned DNA and *in vitro* mutagenesis.

This approach has already been applied to two other viral glycoproteins, to investigate the importance of their cytoplasmic domains for transport to the cell surface, yielding opposite conclusions[40,41]. An intrinsic problem with the method, however, is the difficulty of distinguishing specific effects due to alterations at the site of mutagenesis, from a general structural disruption of the molecule. In this respect the E1 protein may be advantageous in that it provides the possibility of creating a more 'active' phenotype in the mutated molecule: specifically, particular alterations to the protein may result in its transport to the cell surface.

We thank Willy Spaan for communicating results before publication, G. Heisterberg-Moutsis (G. B. F. Braunschweig) for help with oligonucleotide synthesis, Ben van der Zeijst for discussion and Annie Steiner for preparing the manuscript. J.A. was supported by fellowships from the Royal Society and the European Molecular Biology Organization, H.N. by the Deutsche Forschungsgemeinschaft, SFB47 (Virologie), Teilprojekt B3, and P.R. by a short-term EMBO fellowship. Some of these results have been presented in preliminary form elsewhere[24,25].

Received 19 December 1983; accepted 21 March 1984.

1. Siekevitz, P. & Palade, G. F. *J. Biophys. Biochem. Cytol.* **7,** 619–630 (1960).
2. Blobel, G. & Dobberstein, B. *J. Cell Biol.* **67,** 835–851 (1975).
3. Green, J., Griffiths, G., Louvard, D., Quinn, P. & Warren, G. *J. molec. Biol.* **152,** 663–698 (1981).
4. Bergmann, J., Tokuyasu, K. & Singer, S. J. *Proc. natn. Acad. Sci. U.S.A.* **78,** 1746–1750 (1981).
5. Roth, J. & Berger, E. G. *J. Cell Biol.* **93,** 223–229 (1982).
6. Dunphy, W. G., Fries, E., Urbani, L. J. & Rothman, J. E. *Proc. natn. Acad. Sci. U.S.A.* **78,** 7453–7457 (1981).
7. Goldberg, D. E. & Kornfeld, S. *J. biol. Chem.* **258,** 3159–3165 (1983).
8. Quinn, P., Griffiths, G. & Warren, G. *J. Cell Biol.* **96,** 851–856 (1983).
9. Rottier, P., Brandenburg, D., Armstrong, J., van der Zeijst, B. & Warren, G. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
10. Siddell, S., Wege, H. & Ter Meulen, V. *J. gen. Virol.* **64,** 761–776 (1983).
11. Holmes, K. V., Doller, E. W. & Sturman, L. S. *Virology* **115,** 334–344 (1981).
12. Dubois-Dalcq, M. E., Doller, E. W., Haspel, M. W. & Holmes, K. V. *Virology* **119,** 317–331 (1982).
13. Niemann, H. *et al. EMBO J.* **1,** 1499–1504 (1982).
14. Tooze, J., Tooze, S. & Warren, G. *Eur. J. Cell Biol.* (in the press).
15. Leibowitz, J. L., Wilhelmsen, K. C. & Bond, C. W. *Virology* **114,** 39–51 (1981).
16. Lai, M M. C. *et al. J. Virol.* **39,** 823–834 (1981).
17. Cheley, S., Anderson, R., Cupples, M. J., Lee Chan, E. C. M. & Morris, V. L. *Virology* **112,** 596–604 (1981).
18. Spaan, W. J. M., Rottier, P. J. M., Horzinek, M. C. & van der Zeijst, B. A. M. *J. Virol.* **42,** 432–439 (1982).
19. Rottier, P. J. M., Spaan, W. J. M. & van der Zeijst, B. A. M. *J. Virol.* **38,** 20–26 (1982).
20. Leibowitz, J. L., Weiss, S. R., Paavola, E. & Bond, C. W. *J. Virol.* **43,** 905–913 (1982).
21. Lai, M. M. C., Patton, C. D., Baric, R. S. & Stohlman, S. A. *J. Virol.* **46,** 1027–1033 (1983).
22. Spaan, W. *et al. EMBO J.* **2,** 1839–1844 (1983).
23. Armstrong, J., Smeekens, S. & Rottier, P. *Nucleic Acids Res.* **11,** 883–891 (1983).
24. Armstrong, J., Smeekens, S., Rottier, P., Spaan, W. & van der Zeijst, B. A. M. in *Molecular Biology and Pathogenesis of Coronaviruses* (eds Rottier, P. J. M., van der Zeijst, B. A. M., Spaan, W. J. M. & Horzinek, M. C.) (Plenum, New York, in the press).
25. Niemann, H., Heisterberg-Moutsis, G., Geyer, R., Klenk, H.-D. & Wirth, M. in *Molecular Biology and Pathogenesis of Coronaviruses* (eds Rottier, P. J. M., van der Zeijst, B. A. M., Spaan, W. J. M. & Horzinek, M. C.) (Plenum, New York, in the press).
26. Messing, J. & Vieira, J. *Gene* **19,** 269–276 (1982).
27. Dente, L., Cesareni, G. & Cortese, R. *Nucleic Acids Res.* **11,** 1645–1655 (1983).
28. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. *J. molec. Biol.* **143,** 161–178 (1980).
29. Skinner, M. A. & Siddell, S. G. *Nucleic Acids Res.* **11,** 5045–5054 (1983).
30. Lai, M. M. C., Patton, C. D. & Stohlman, S, A. *J. Virol.* **41,** 557–565 (1982).
31. Lai, M. M. C., Patton, C. D. & Stohlman, S. A. *J. Virol.* **44,** 487–492 (1982).
32. Jacobs, L., Spaan, W. J. M., Horzinek, M. C. & van der Zeijst, B. A. M. *J. Virol.* **39,** 401–406 (1981).
33. Sturman, L. S. *Virology* **77,** 637–649 (1977).
34. Von Heijne, G. *Eur. J. Biochem.* **133,** 17–21 (1983).
35. Sturman, L. S. & Holmes, K. V. *Virology* **77,** 650–660 (1977).
36. Niemann, H. & Klenk, H.-D. *J. molec. Biol.* **153,** 993–1010 (1981).
37. Housman, D., Jacobs-Lorena, M., Rajbhandary, U. L. & Lodish, H. F. *Nature* **227,** 913–918 (1970).
38. Furthmayr, H. *Nature* **271,** 519–524 (1978).
39. Niemann, H. *et al. EMBO J.* **3,** 665–670 (1984).
40. Garoff, H., Kondor-Koch, C., Pettersson, R. & Burke, B. *J. Cell Biol.* **97,** 652–658 (1983).
41. Rose, J. K. & Bergmann, J. E. *Cell* **34,** 513–524 (1983).