CARBOXYL-TERMINAL SEQUENCE ANALYSIS OF THE FOUR STRUCTURAL PROTEINS OF SEMLIKI FOREST VIRUS

Nisse KALKKINEN

Department of Biochemistry, University of Helsinki, Unioninkatu 35, SF-00170 Helsinki 17, Finland

Received 28 April 1980

1. Introduction

Semliki Forest virus, an extensively studied alphavirus, contains 4 different structural proteins. The capsid protein (M_r 33 000) together with the RNA genome forms the nucleocapsid. The 3 envelope glycoproteins E1 (M_r 49 000), E2 (M_r 52 000) and E3 (M_r 10 000) are constituents of the lipoprotein membrane (envelope) which surrounds the nucleocapsid [1,2]. E1 and E2 are integral membrane proteins, whereas E3 does not interact with the membrane lipids [3–5].

The structural proteins are translated essentially as a polyprotein from a virus specific subgenomic messenger RNA (26 S RNA). The individual proteins are then formed from this precursor by post-translational cleavage (reviewed [1,2]). The order of the genes in the 26 S RNA is as follows: capsid protein–E3–E2–E1. The N-terminal capsid protein is translated on free ribosomes and is then cleaved from the growing polyprotein. This possesses a signal sequence of the same kind to those of secretory proteins [6–8] which results in the binding of the ribosomes to the endoplasmic reticulum membrane and the transfer of the following glycoprotein p62 (precursor for E3 and E2) across the membrane. A second signal sequence responsible for the transfer of E1 protein has also been proposed (L. Kääriäinen, personal communication).

The structural proteins are formed by proteolytic processing of the primary translation product. However, nothing is known about the specificity of the cleaving enzymes. In this context the structures around the cleavage sites are of considerable interest.

Knowledge of the N- and C-terminal amino acid sequences of the structural proteins is vital for the interpretation of the nucleotide sequence of the 26 S RNA, which is soon determined (K. Simons and H. Garoff, personal communication). When this information is available, the important amino acid sequences (e.g., lead in sequences, signal sequences) which are not present in the final processed translation products can be elucidated from the nucleotide sequence.

The structural proteins have previously been subjected to N-terminal amino acid sequence analysis [9]. About 20 residues have been determined for both the E1 and E2 proteins. The capsid and E3 proteins were found to have blocked N-termini. Here, the results of C-terminal amino acid sequence analysis of subnanomolar amounts of the 4 structural proteins are presented.

2. Materials and methods

The four structural proteins of Semliki Forest virus were purified as in [9]. C-terminal amino acid sequences were determined by digestion of the proteins with carboxypeptidase A (CPA) and carboxypeptidase B (CPB). CPA solution was prepared by adding 1 μl diisopropylfluorophosphate-CPA suspension (24 mg/ml, 45 U/mg, Sigma, St Louis) to 50 μl 50 mM $NaHCO_3$ (pH 8.15). Then 25 μl 1 N NaOH was added to dissolve the protein. The pH was brought back to 8–9 with 25 μl 0.1 N HCl and made up to 250 μl with 50 mM $NaHCO_3$ (pH 8.15). Before use, diisopropylfluorophosphate-CPB (2 mg/ml, 69 U/mg, Sigma, St Louis) was further purified by gel filtration on a Sephadex G-25 Fine (0.7 × 13 cm, 0.1 M NaCl) column. The CPA and CPB activities were determined using hippuryl-L-phenylalanine and hippuryl-L-arginine as substrates, respectively [10,11]. CPA was found to be free from CPB activity whereas the CPB preparation contained 3.3% (in units) CPA.

For qualitative analyses 0.2–0.4 nmol protein in 25 μl 50 mM $NaHCO_3$ (pH 8.15), 0.05% sodium dodecyl sulphate (SDS) was digested with CPA, CPB or both. The reactions were terminated by rapid freezing followed by lyophilization. For dansylation, the lyophilized samples were dissolved in 20 μl 0.1 M $NaHCO_3$ (pH 9.2) and 10 μl dansyl chloride (5 mg/ml acetone) was added. The reaction time was 15 min at 50°C followed by drying in reduced pressure at 50°C. The dansylated amino acids were identified by thin layer chromatography on 3 X 3 cm polyamide sheets (Cheng Ching Trading Co., Taiwan) [12].

For quantitative analyses 0.6–0.8 nmol protein was digested as above. The released amino acids were then analyzed on a Beckman 121M automatic amino acid analyzer. Protein amounts were determined by quantitative amino acid analysis after 18 h hydrolysis in 6 N HCl at 110°C.

For enzyme blanks the highest used amounts of CPA (14 pmol) and/or CPB (18 pmol) were incubated without substrate for 2 h at 37°C. For protein blanks 0.79 nmol E1, 0.56 nmol E2, 0.67 nmol E3 and 0.76 nmol capsid protein were analyzed. No significant amounts of free amino acids could be detected when the enzyme and protein blanks were subjected to quanitative and qualitative amino acid analysis.

3. Results

The four structural proteins of Semliki Forest virus were isolated and digested with CPA and CPB. Because only limited amounts of the proteins were available, optimal digestion conditions for the quantitative analyses were first determined on the basis of qualitative experiments, using the more sensitive dansylation/thin layer chromatography method.

E1 protein was digested as shown in fig.1A. Digestion with CPB for 1 h liberated 1.71 mol arginine and 0.24 mol leucine/mol protein. Addition of CPA to this mixture and an additional incubation for 1 h resulted in a total liberation of 1.96 mol arginine and 0.72 mol leucine/mol protein. According to this data, the C-terminal amino acid sequence of E1 protein is –Leu–Arg–Arg. This structure was also supported by the following qualitative results. When E1 protein was first incubated with CPA (57:1, mol substrate:mol enzyme) for 2 h at 37°C, no amino acids were released. Incubation of 0.39 nmol E1 protein with CPB in milder conditions (265:1, mol substrate:mol enzyme, 30°C, 20 min) released only arginine in detectable amounts.

E2 protein was digested as shown in fig.1B. Digestion with CPA for 1 h released 2.01 mol alanine and 0.92 mol histidine mol protein. Addition of CPB to this mixture and an additional incubation for 1 h resulted in a total release of 2.01 mol alanine, 1.06 mol histidine and 0.60 mol arginine/mol protein. As can be seen, no alanine was released during the second hour of incubation, whereas still an additional amount of 0.14 mol histidine and 0.60 mol arginine were released/mol protein. Thus the C-terminal amino acid sequence of E2 protein is probably –Arg–His–Ala–Ala. The C-terminal alanine could be qualitatively detected already in very mild digestion conditions when 0.28 nmol E2 was digested with CPA (3650:1, mol substrate:mol enzyme) for 15 min at 20°C.

E3 protein was digested as shown in fig.1C. Digestion with CPB for 1 h released 0.92 mol arginine and 0.69 mol histidine/mol protein. An additional incubation of this mixture with CPA for 1 h resulted in a total liberation of 1.11 mol arginine and 0.86 mol histidine/mol protein. Additional qualitative results supporting a C-terminal amino acid sequence –His–Arg for E3 protein were also obtained. When E3 was incubated with CPA (26:1, mol substrate:mol enzyme) for 45 min at 30°C no released amino acids could be detected. Digestion of E3 in milder conditions with CPB (168:1, mol substrate:mol enzyme, 30 min, 20°C) released only arginine in detectable amounts.

Incubation of the capsid protein with CPA resulted in the liberation of 1.05 mol tryptophan/mol protein and an additional digestion with CPB did not result in an additional release of amino acids as shown in fig.1D.

The release of leucine from E1 and histidine from E3 with the CPB preparation can be explained by its contaminating CPA activity.

The obtained C-terminal structures of the Semliki Forest virus structural proteins are summarized in table 1.

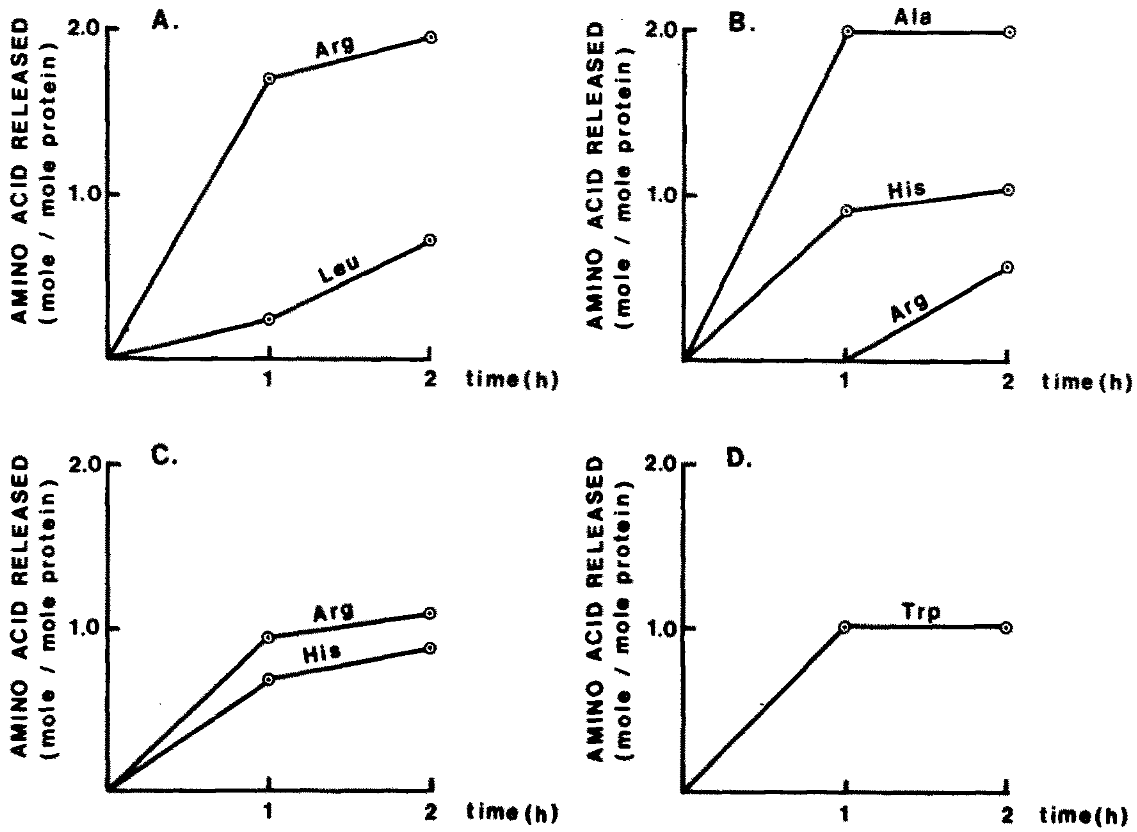

Fig.1. Amino acids released by carboxypeptidase (CP) digestion of the 4 structural proteins of Semliki Forest virus in 50 mM $NaHCO_3$ (pH 8.15), 0.05% SDS at 37°C. The substrate to enzyme ratios are expressed as mol substrate:mol enzyme. (A) 0.79 nmol E1 protein digested with CPB (44:1) for 1 h followed by addition of CPA (57:1) and a further incubation for 1 h; (B) 0.56 nmol E2 protein digested with CPA (40:1) for 1 h followed by addition of CPB (31:1) and a further incubation for 1 h; (C) 0.67 nmol E3 protein digested with CPB (37:1) for 1 h followed by addition of CPA (48:1) and an additional incubation for 1 h. (D) 0.76 nmol capsid protein digested with CPA (27:1) for 1 h followed by addition of CPB (21:1) and an additional incubation for 1 h.

4. Discussion

Precise location of the genome regions coding for the Semliki Forest virus structural proteins is impossible without knowledge of the N- and C-terminal amino acid sequences of the structural proteins.

These results should be sufficient to localize the nucleotide sequences on the genome coding for the C-termini of the structural proteins. For E1 and E2 proteins, the sequence of 3 and 4 amino acids, respectively, determined, should be enough to locate unambiguously the corresponding nucleotide sequences. The sequence of the two C-terminal amino acids of E3 protein is probably also sufficient, especially since the reading frame is known from the N-terminus of the E2 protein. When the capsid protein was digested with CPA and CPB, only tryptophan was detected indicating that the next amino acid most probably is one of those not at all or only very slowly released with a mixture of CPA and CPB (Pro, Gly, Asp, Glu) [13]. Tryptophan is not a common amino acid in proteins [14] and is coded for by only one codon, UGG. Thus, it seems most likely that also the region in the genome coding for the C-terminus of the capsid protein can be identified.

Table 1
Carboxyl-terminal structures of the 4 structural proteins of Semliki Forest virus

E1 protein	–Leu–Arg–Arg
E2 protein	–Arg–His –Ala–Ala
E3 protein	–His –Arg
Capsid protein	–Trp

The structural proteins are formed by proteolytic processing. The C-terminal structures obtained do not show any sequence homology to each other so argueing against a common substrate specificity of the cleaving enzymes. The C-terminus of the capsid protein is formed in the cytoplasm, the C-terminus of E3 at the cell plasma membrane and the C-terminus of E2 at the endoplasmic reticulum membrane. It is more likely that these cleavages, occurring in different compartments of the cell, require proteases of different substrate specificity.

The E1 and E2 proteins are anchored to the lipid bilayer with hydrophobic segments locating in their C-terminal regions and at least E2 protein spans the lipid bilayer [4]. The charged C-terminal amino acid sequence of E1 (–Arg–Arg) demonstrated here suggests that also this C-terminus must be located outside the hydrophobic interior of the membrane.

Acknowledgements

I would like to thank Dr Hans Jörnvall for the possibility to use the Beckman 121M amino acid analyzer and Dr Leevi Kääriäinen, Dr Hans Söderlund and Dr Päivi Lehtovaara for helpful discussions and constructive criticism of the manuscript. This investigation was partly supported by the Emil Aaltonen Foundation, Oskar Öflund Foundation and Lääke Oy Research Foundation.

References

[1] Strauss, J. H. and Strauss, E. G. (1977) in: The molecular biology of animal viruses (Nyak, D. P. ed) pp. 111–116, Dekker, New York.

[2] Kääriäinen, L. and Söderlund, H. (1978) Curr. Top. Microbiol. Immunol. 82, 15–69.

[3] Uterman, G. and Simons, K. (1974) J. Mol. Biol. 85, 569–587.

[4] Garoff, H. and Söderlund, H. (1978) J. Mol. Biol. 124, 535–549.

[5] Zimiecki, A. and Garoff, H. (1978) J. Mol. Biol. 122, 259–269.

[6] Campbell, P. N. and Blobel, G. (1976) FEBS Lett. 72, 215–226.

[7] Rothman, J. E. and Lodish, H. F. (1977) Nature 269, 775–780.

[8] Irving, R. A., Toneguzzo, F., Rhee, S. H., Hofman, T. and Ghosh, H. P. (1979) Proc. Natl. Acad. Sci. USA 76, 570–574.

[9] Kalkkinen, N., Jörnvall, H., Söderlund, H. and Kääriäinen, L. (1980 Eur. J. Biochem. in press.

[10] Folk, J. E. and Schirmer, E. W. (1963) J. Biol. Chem. 238, 3884–3894.

[11] Folk, J. E., Piez, K. A., Carrol, W. R. and Gladner, J. A. (1960) J. Biol. Chem. 235, 2272–2277.

[12] Weiner, A. M., Platt, T. and Weber, K. (1972) J. Biol. Chem. 247, 3242–3251.

[13] Ambler, R. P. (1972) Methods Enzymol. 25, 262–272.

[14] Dayhoff, M. O., Hunt, L. T. and Hurst-Calderone, S. (1978) in: Atlas of Protein Sequence and Structure (Dayhoff, M. O. ed) vol. 5, suppl. 3, pp. 363–375, National Biomedical Research Foundation, Washington.