J. Mol. Biol. (1981) **152**, 171–179

Amino Acid Sequence of the Coat Protein Subunit in Satellite Tobacco Necrosis Virus

The primary structure of the coat protein subunit in satellite tobacco necrosis virus has been investigated. The results obtained are consistent with and support the proposal for the amino acid sequence made from the nucleotide sequence of RNA (Ysebaert *et al.*, 1980). This would imply that no intervening sequences of RNA occur in the cistron for the satellite tobacco necrosis virus coat protein. The polypeptide chain of the protein consists of 195 amino acid residues. It contains one sulfhydryl group but no disulfide bridges. The distribution of various kinds of amino acid residues along the chain is markedly uneven.

The present investigation is part of a joint effort by several laboratories to determine the complete structure of the satellite tobacco necrosis virus. X-ray diffraction studies on STNV† are being carried out by B. Strandberg and his collaborators in Uppsala. An electron density distribution to a resolution of 4·0 Å has recently been completed (Unge *et al.*, 1980). A detailed interpretation of the electron density maps from the X-ray studies would require knowledge about the covalent structures of the nucleic acid of the virus, and of the protein constituting the virus coat. Determination of the structure of the nucleic acid is described in two previous papers (van Emmelo *et al.*, 1980; Ysebaert *et al.*, 1980), and the coat protein sequence is dealt with in the present study.

STNV is a small plant virus that for its multiplication is dependent upon the presence of a helper virus, the tobacco necrosis virus (Kassanis & Nixon, 1961) (for a discussion on satellitism among plant viruses, see Atabekov, 1977). The simplicity of STNV makes it an attractive object in studies of structure and function of viruses. It is one of the smallest virus particles known, having a molecular weight of about $1{\cdot}7 \times 10^6$ (Sjöberg, 1977). It has a coat of 60 identical protein subunits of molecular weight 21,600 forming an icosahedron (Unge *et al.*, 1980). Inside the coat is a single-stranded RNA that is believed to be monocistronic with the coat protein subunit as the only polypeptide species coded for (Leung *et al.*, 1976).

The studies by Leung *et al.* (1979) and Ysebaert *et al.* (1980) have shown that the structural gene for the coat protein starts close to one end of the nucleic acid molecule, with the translation initiation codon AUG at positions 30–32 from the 5′ terminus of the RNA strand. The RNA molecule consists of 1239 nucleotides (Ysebaert *et al.*, 1980) which are about twice the number of nucleotides required to code for the coat protein subunit. This would allow the possibility for intervening sequences in the RNA strand that do not become translated during protein synthesis. To settle this question would require a comparison of the nucleic acid sequence with that of the protein. The present study provides partial information about the protein sequence to be used for this and other purposes. It has been

† Abbreviation used: STNV, satellite tobacco necrosis virus.

0022–2836/81/290171–09 $02.00/0

obtained by isolation and characterization of fragments derived from the protein by hydrolysis with trypsin or thermolysin, or after cleavage with cyanogen bromide.

Whole STNV particles, prepared as described by Fridborg *et al.* (1965), were obtained from Dr B. Strandberg, Uppsala. The coat protein was purified from the virus particles essentially as described by Rees *et al.* (1970). The viral RNA was hydrolyzed by suspending the intact virus in 1 M-HCl for 18 hours at room temperature. The solution was then neutralized with 1 M-NaOH and dialyzed against distilled water until free from salt.

L-(Tosyl 2-phenyl)ethyl chloromethyl ketone-treated trypsin and diisopropyl fluorophosphate-treated carboxypeptidase A were purchased from Worthington Biochemical Corporation, and thermolysin from Boehringer Mannheim Biochemicals. Cyanogen bromide, guanidine·hydrochloride and reagents for automatic sequencing were purchased from Pierce Chemical Co., and Sephadex gels from Pharmacia Fine Chemicals. All other chemicals were of reagent grade quality.

Cyanogen bromide cleavage was performed on 40 mg of STNV coat protein. The protein was dissolved in 5 ml of 90% (v/v) formic acid, and 0·2 g of cyanogen bromide was added, giving a cyanogen bromide/Met ratio of 200 : 1. After 20 hours at room temperature the formic acid was diluted to 40% and the reaction mixture lyophilized. The material was then dissolved in 1 ml of saturated guanidinium chloride in 1 M-acetic acid and applied to a 1 cm × 120 cm column of Sephadex G50 eluted with 5 M-guanidinium chloride (pH 5). The fractions collected were read at an absorbance of 280 nm and 230 nm.

Tryptic digestion was performed on 50 mg of coat protein. The protein suspension was titrated to pH 10·6 and 1 mg of trypsin was added. After one hour at 37°C 1 ml of 0·2 M-ammonium bicarbonate and 0·5 mg of trypsin were added. After 18 hours at 37°C the insoluble peptides were removed by centrifugation. The supernatant was concentrated under a stream of nitrogen, and applied to a 1 cm × 120 cm Sephadex G25F column eluted with 0·15 M-*N*-ethyl morpholine (pH 9·0). A portion (1/100) of each fraction was spotted on paper, separated by high-voltage electrophoresis (Smillie & Hartley, 1966) at pH 6·5 and stained with Cd ninhydrin (Easley, 1965). Fractions were pooled on the basis of the Cd ninhydrin staining. Peptides in the pooled fractions were separated by paper electrophoresis and by paper chromatography in butanol/acetic acid/water (Light & Smith, 1962). The insoluble tryptic peptides were extracted with saturated guanidinium chloride in 50% formic acid. Material soluble under these conditions was separated on a 1 cm × 120 cm Sephadex G50F column eluted with 5 M-guanidinium chloride (pH 5).

Tryptic digestion was also performed on performic acid-oxidized protein and the peptide containing the cysteic acid residue was purified by gel filtration on Sephadex G50F in 1 M-acetic acid and by paper electrophoresis.

In an attempt to produce overlap peptides 10 mg of coat protein were digested with thermolysin. Digestion conditions were: 0·2 M-ammonium bicarbonate, 50 : 1 protein to enzyme ratio for 20 hours at 55°C. The resulting peptide mixture was separated by gel filtration on a Sephadex G25F column in 0·15 M-*N*-ethyl morpholine (pH 9·0).

Amino acid analyses were performed on samples hydrolyzed *in vacuo* in 6 M-HCl at 105°C for 18 hours. A one-column system (Benson, 1974) with a Hamilton H-70 resin was used.

Sequencing of peptides and of the amino terminus of the coat protein was carried out by Edman degradation using a liquid-phase Beckman model 890 B sequencer. Phenylthiohydantoin derivatives of amino acids were identified by thin-layer chromatography (Summers *et al.*, 1973) and gas chromatography on SP-400 or SE-30 coated supports. Through the courtesy of Drs M. Hunkapiller and L. E. Hood, California Institute of Technology, the amino terminus of the coat protein was also investigated using their home-built spinning-cup sequencer (Hunkapiller & Hood, 1978).

The molecular weight of the STNV coat protein was investigated by gel electrophoresis in the presence of sodium dodecyl sulphate (Weber & Osborn, 1975).

The results of the present investigation are summarized in Figure 1 and Table 1. Table 1 gives the amino acid compositions of the various fragments derived from the coat protein by the action of trypsin, thermolysin, or cyanogen bromide. The amino acid composition of the whole protein has also been included in Table 1. The figures represent the data of Unge & Strandberg (1979), which are in close accordance with earlier determinations by Reichman (1964) and Rees *et al.* (1970). The composition of the protein determined by amino acid analysis is in good agreement with the theoretical values estimated from the amino acid sequence. The only noticeable deviation is for glycine where the value from the acid hydrolysate is higher than the theoretical one.

Figure 1 gives a survey of the sequencing work carried out on the coat protein. The amino terminus of the protein was investigated by Edman degradation. The first 60 steps allowed distinct assignments. The protein was hydrolyzed with trypsin. From the digest 14 peptides, together comprising 143 residues, were purified and their amino acid compositions are in Table 1. All of these were completely sequenced except for peptides 3–8, 130–143, and 146–195. The latter was sequenced from the amino terminus through residue 173. The tryptic digest was also found to contain free lysine and arginine indicating the occurrence of clusters of basic amino acid residues in the protein sequence.

Serious problems were encountered in the purification of several of the tryptic peptides. These problems can be ascribed to the occurrence of long hydrophobic stretches in the protein sequence (see for instance the region 146–177 in Fig. 1, where out of 32 residues none has an ionizable side-chain) and an uneven distribution of lysine and arginine residues. As an example, the two tryptic peptides 97–123 and 146–195 were extracted in rather low yields from the tryptic core by a saturated solution of guanidinium chloride in 50% acetic acid. When separated on a Sephadex column in 5 M-guanidinium chloride they emerged from the column as precipitates, which facilitated the removal of the guanidinium chloride. Tryptic peptide 29–60 was not extracted under the above conditions and its sequence was seen together with those of peptides 97–123 and 146–195 when the extracted core material was subjected to sequential degradation.

In order to obtain overlap information for the tryptic peptides the protein was digested with thermolysin. Due to the high number of cleavage points only four

TABLE 1

Amino acid composition of STNV coat protein and of peptides derived from the protein after treatment with trypsin, thermolysin, or cyanogen bromide

Amino acid residue	1–2 (T)	3–8 (T)	10–14 (T)	15–17 (T)	19–27 (T)	61–66 (T)
Asx		2·0(2)			1·1(1)	1·2(1)
Thr			1·0(1)		1·1(1)	
Ser			1·1(1)			0·8(1)
Glx		2·0(2)			1·1(1)	1·2(1)
Pro						
Gly						1·0(1)
Ala	1·0(1)		1·1(1)	1·0(1)		
Cys						
Val				1·0(1)		0·8(1)
Met			1·0(1)		0·9(1)	
Ile					1·0(1)	
Leu					1·0(1)	
Tyr						
Phe						
His					2·0(2)	
Lys	1·0(1)			1·2(1)	1·0(1)	
Arg		1·7(2)	1·0(1)			1·0(1)
Trp†						
Total	2	6	5	3	9	6

Amino acid residue	67–71 (T)	72–75 (T)	92–96 (T)	97–123 (T)	125–129 (T)	130–143 (T)
Asx			3·0(3)	3·1(3)		2·1(2)
Thr				5·4(5)	1·1(1)	1·8(2)
Ser	1·0(1)			0·7(1)		1·6(2)
Glx				4·3(4)		1·2(1)
Pro				2·5(2)		
Gly				0·9(1)		1·1(1)
Ala				0·9(1)		0·2(0)
Cys						1·0‡(1)
Val	0·5(1)	1·1(1)		1·5(2)		1·0(1)
Met			1·0(1)	0·7(1)		
Ile	0·5(1)			1·0(1)	1·0(1)	0·9(1)
Leu		1·0(1)		3·0(3)	1·0(1)	1·8(2)
Tyr				0·8(1)		
Phe				0·7(1)	0·9(1)	
His	1·0(1)	0·9(1)				
Lys	1·1(1)			0·9(1)	1·3(1)	1·1(1)
Arg		1·0(1)	1·4(1)			
Trp†						
Total	5	4	5	27	5	14

Amino acid residue	144–145 (T)	146–195 (T)	85–86 (Tl)	88–89 (Tl)	120–124 (Tl)	127–130 (Tl)
Asx	1·1(1)	6(7)			1·1(1)	1·1(1)
Thr		2(2)				
Ser		4(4)				
Glx		3(3)			1·9(2)	0·3(0)
Pro		2(2)				
Gly		6(6)				
Ala		6(7)				
Cys						
Val		2(4)				
Met		1(1)				
Ile		3(4)		+(1)		0·9(1)
Leu		4(5)				1·0(1)
Tyr		2(3)				
Phe		1(1)	1·0(1)			
His						
Lys					1·1(1)	1·1(1)
Arg	1·0(1)		1·0(1)		0·9(1)	0·2(0)
Trp†		+(1)		+(1)		
Total	2	50	2	2	5	4

Amino acid residue	95–173 (CNBr)	174–195 (CNBr)	STNV coat protein
Asx	8·8(11)	2·8(3)	28·0(28)
Thr	7·1 (9)	0·9(1)	18·0(18)
Ser	5·5 (4)	2·6(3)	13·0(13)
Glx	7·0 (6)	2·0(2)	16·4(16)
Pro	n.d. (4)		3·9 (4)
Gly	6·3 (6)	2·0(2)	17·5(14)
Ala	5·9 (6)	2·0(2)	14·7(15)
Cys	— (1)		1·0‡(1)
Val	5·8 (5)	2·0(2)	12·6(13)
Met	0 (2)		5·3§ (5)
Ile	4·1 (6)	1·1(1)	13·0(15)
Leu	7·0 (8)	2·7(3)	15·2(15)
Tyr	3·7 (2)	1·8(2)	3·8 (4)
Phe	4·0 (3)		6·5 (7)
His			3·8 (4)
Lys	3·9 (3)		8·0 (8)
Arg	3·4 (3)		13·0(13)
Trp†		+(1)	n.d. (2)
Total	79	22	195

The peptide numbers refer to the positions of the first and last residues of the peptide in the coat protein sequence (see Fig. 1). The letters within brackets indicate the type of cleavage used to produce the peptide: trypsin (T), thermolysin (Tl), cyanogen bromide (CNBr). Amino acid compositions are given as the number of residues per molecule. The figures within parentheses are the theoretical values deduced from the amino acid sequence.

† Detected by Ehrlich stain (Easley, 1965).

‡ Determined as cysteic acid.

§ Determined as methionine sulfone.

n.d., not determined.

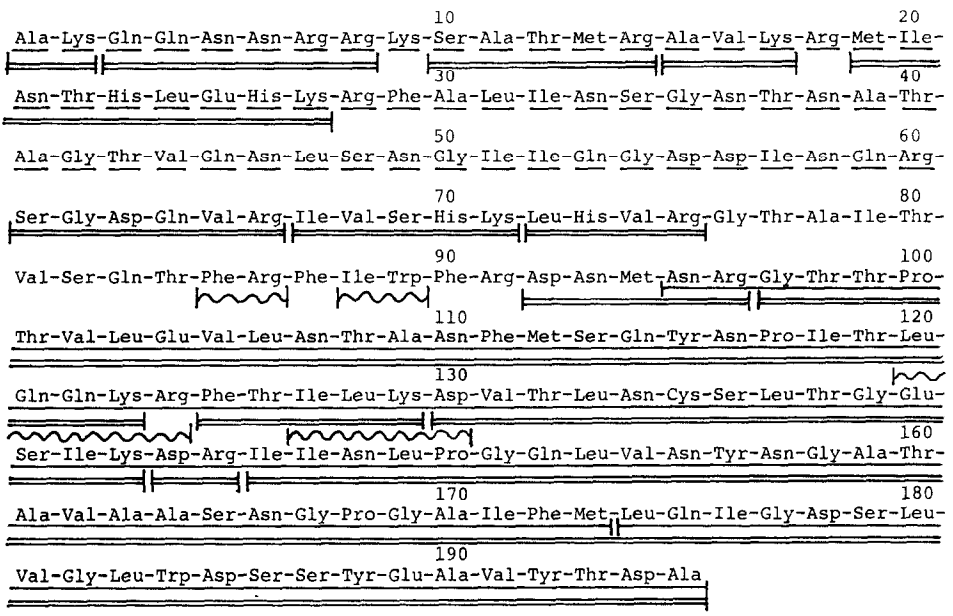

FIG. 1. Primary structure of the STNV coat protein. (The complete amino acid sequence is that deduced from the nucleic acid sequence (Ysebaert *et al.*, 1980).) The amino terminal sequence of the protein was investigated by Edman degradation (– – –) using a liquid-phase sequencer. Of the tryptic peptides isolated (|═══|) all were completely sequenced except for peptides 3–8 and 130–143, and peptide 146–195 that was sequenced through residue 173. Of the cyanogen bromide fragments (|———|), 95–173 was subjected to 17 steps of Edman degradation while the carboxyl terminal one (174–195) was completely sequenced as were the thermolytic peptides (|∼∼∼|). The amino acid compositions of the peptides are given in Table 1 together with the composition of the whole protein.

peptides were completely purified (see Fig. 1). The thermolytic digest was separated by gel filtration on Sephadex G25F in 0·15 M-*N*-ethyl morpholine. Under these conditions thermolytic peptides 88–89 (Ile-Trp) and 85–86 (Phe-Arg) appeared as strongly retarded.

After treatment of the coat protein with cyanogen bromide two fractions (see Table 1) were isolated. The first of these (95–173) was found to consist of fragment 95–173 contaminated with fragment 113–173. This was shown by 17 steps of Edman degradation that yielded double answers on each step. The results could be interpreted using the previously determined sequences of tryptic peptides 92–96, 97–123, and 125–129. Obviously, the Met–Ser peptide bond at position 112–113 is only partially cleaved with cyanogen bromide. Difficulties to cleave Met–Thr and Met–Ser peptide bonds with this reagent have been reported (e.g. see Chen *et al.*, 1975). The second fraction isolated from the cyanogen bromide-treated coat protein (174–195) was found to be homogeneous and was completely sequenced.

The molecular weight for the coat protein determined by gel electrophoresis in the presence of sodium dodecyl sulphate was found to be 22,200.

The sequence information for the STNV coat protein presented here is not

sufficient for a complete and independent determination of the primary structure of the protein. When the information about the nucleic acid sequence became available to us (W. Fiers, personal communication) we found a good agreement with our partial data for the amino acid sequence. We decided then that it would not be justified to spend efforts making the protein sequence determination complete in all its details. We believe that the information gathered up to now (see Fig. 1), together with the knowledge of the nucleic acid sequence (Ysebaert *et al.*, 1980), will give a proposal for the primary structure of the coat protein to a level of confidence sufficient for most purposes. For a judgement of this statement we will try below to specify the limitations of our results.

All amino acid residues except 76–91 were found either as tryptic peptides or through direct sequencing of the protein. This region is highly hydrophobic and if the arginyl bond at residue 86 is not hydrolyzed with trypsin the resulting peptide (76–91) would probably be strongly adsorbed to the gel column used for separation of the tryptic peptides. However, some information from this region has been obtained through the two small peptides, 85–86 (Phe-Arg) and 88–89 (Ile-Trp), isolated from the thermolytic digest of the whole protein.

The amino terminus of the coat protein was investigated by Edman degradation. A total of 60 steps of unambiguous assignments were obtained. In the next ten steps the identifications were tentative but in good agreement with the sequences of peptides 61–66 followed by 67–71. In this way the tryptic peptides of the amino terminal region through residue 71 have been arranged in order. However, no evidence has been obtained for the ordering of tryptic peptides 72–75, 76–86, and 87–91.

The region 92–195 is completely accounted for by tryptic peptides which can be ordered using the sequence information from cyanogen bromide and thermolytic peptides derived from this region. The information missing in this region is the sequence of tryptic peptide 130–143. This peptide was purified at about the time when the corresponding region of the nucleic acid was sequenced, and as the composition of the peptide (Table 1) was found to fit completely the nucleic acid sequence it was not investigated further.

Previous studies on the primary structure of the STNV coat protein are consistent with the results obtained here. Leung *et al.* (1979) reported the sequence for the first seven positions of the amino terminal end. Tryptic peptides from the coat protein were purified by Rees *et al.* (1970), who published the amino acid compositions for a large number of peptides. Their results are in quite good agreement with our data.

The results we have obtained are also in good agreement with the results of the nucleic acid sequence determination (Ysebaert *et al.*, 1980). The only position where a difference exists is residue 113. Due to a high background the identification of phenylthiohydantoin threonine from this position was, however, somewhat ambiguous. Furthermore, the amino acid composition of peptide 97–123 used for sequencing of this region shows five residues of threonine and one of serine (see Table 1), while the Edman degradation yielded six threonine residues, whereas no serine was observed. Considering the amino acid composition and the normally low yield of phenylthiohydantoin serine (Horn & Bonner, 1977) we are not inclined to

dispute the serine residue assigned to position 113 from the nucleic acid sequence. At position 166 we were unable to identify any residue, and at position 171 the distinction between leucine and isoleucine could not be made. All other residues being sequenced on peptides or on the whole protein agree with the amino acid sequence deducable from the nucleic acid sequence.

The finding that our results on the protein agree well with the sequence of the cistronic region of the nucleic acid provides an independent support for the nucleotide sequence published by Ysebaert *et al.* (1980). A comparison of the protein sequence information with that for the nucleic acid also provides evidence for the absence of intervening stretches of nucleotides in the RNA molecule that do not become translated into protein structure.

The STNV coat protein consists of 195 amino acid residues in one polypeptide chain without disulfide bridges. It contains a single sulfhydryl group and has the following amino acid composition: Asn19, Asp9, Thr18, Ser13, Gln12, Glu4, Pro4, Gly14, Ala15, Val13, Cys1, Met5, Ile15, Leu15, Tyr4, Phe7, Trp2, Lys8, His4, Arg13. The molecular weight estimated from this composition is about 21,600, in good agreement with the value of 22,200 we obtained by gel electrophoresis in the presence of sodium dodecyl sulphate.

The distribution of various kinds of amino acid residues along the sequence is very uneven. There are long stretches without any ionizable amino acid side-chains and hydrophobicity may play a role in the protein–protein interaction as well as for the conformation of the individual protein subunit. The basic amino acid residues are concentrated to the amino terminal portion of the protein. Among the first 28 residues in the amino terminal region there are 11 basic ones. The interior of the polypeptide chain is largely hydrophobic whereas the carboxyl terminal region is slightly acidic. The distribution of residues would suggest that the amino terminal part of the protein is located on the inside of the protein shell, binding to the negatively charged nucleic acid.

A high frequency of basic residues in the amino terminus of the coat protein has also been found for several other plant viruses (see Argos, 1980), suggesting a similarity in function of this region in different species of virus. Predictions of secondary structure from coat protein sequences of STNV and other plant viruses (Argos, 1980) suggest that the amino terminal region has a helical conformation. This is consistent with the X-ray studies on STNV (Unge *et al.*, 1980) which show that the coat protein subunit extends with an arm interpreted as a helix into the interior of the virus particle. This helical portion was identified as the amino terminus of the protein, since it could be labelled with iodine which is known to become covalently attached to His23 upon iodination of the STNV particle (Unge & Strandberg, 1979).

We are indebted to Dr B. Strandberg and his collaborators for providing us with STNV for this investigation. We also thank these people as well as Professor W. Fiers and his collaborators for submitting to us their results on STNV prior to publication. We are very grateful to Drs M. Hunkapiller and L. E. Hood at the California Institute of Technology for analyzing a sample of the STNV coat protein in their home-built spinning-cup sequencer. This investigation has been aided by grants from the Swedish Natural Science Research

Council, Magnus Bergvalls Stiftelse, Wilhelm och Martina Lundgrens Vetenskapsfond, and United States Public Health Service grants GM–15419 and GM–24681.

[1] Department of Human Genetics
University of Michigan Medical School
Ann Arbor, Mich. 48109, U.S.A.

D. HENRIKSSON[1,2]
R. J. TANIS[1]
R. E. TASHIAN[1]
P. O. NYMAN[2]

[2] Department of Biochemistry and Biophysics
University of Göteborg and Chalmers Institute of Technology
S-412 96 Gothenburg, Sweden

Received 3 October 1980

REFERENCES

Argos, P. (1980). In *Proceedings of the 7th Aharon Katzir-Katchalsky Conference on Structural Aspects of Recognition and Assembly in Biological Macromolecules*, Nof Ginossor, Israel, in the press.

Atabekov, J. G. (1977). *Comprehensive Virology* (Fraenkel-Conrat, H. & Wagner, R. R., eds), vol. 11, pp. 143–200, Plenum Press, New York.

Benson, J. V. (1974). *Analysis*, vol. 1, no. 2, Hamilton Company, Reno.

Chen, K. C. S., Tao, N. & Tang, J. (1975). *J. Biol. Chem.* **250**, 5068–5075.

Easley, C. W. (1965). *Biochim. Biophys. Acta*, **107**, 386–388.

Fridborg, K., Hjertén, S., Höglund, S., Liljas, A., Lundberg, B. K. S., Oxelfelt, P., Philipson, L. & Strandberg, B. (1965). *Proc. Nat. Acad. Sci., U.S.A.* **54**, 513–521.

Horn, M. J. & Bonner, A. G. (1977). *Solid Phase Methods in Protein Sequence Analysis* (Previero, A. & Coletti-Previero, M.-A., eds), pp. 163–176, North-Holland Publishing Company, Amsterdam.

Hunkapiller, M. W. & Hood, L. E. (1978). *Biochemistry*, **17**, 2124–2133.

Kassanis, B. & Nixon, H. L. (1961). *J. Gen. Microbiol.* **25**, 459–471.

Leung, D. W., Gilbert, C. W., Smith, R. E., Sasavage, N. L. & Clark, J. M. (1976). *Biochemistry*, **15**, 4943–4950.

Leung, D. W., Browning, K. S., Heckman, J. E., Raj Bhandary, U. L. & Clark, J. M. (1979). *Biochemistry*, **18**, 1361–1366.

Light, A. & Smith, E. L. (1962). *J. Biol. Chem.* **237**, 2537–2546.

Rees, M. W., Short, M. N. & Kassanis, B. (1970). *Virology*, **40**, 448–461.

Reichman, M. E. (1964). *Proc. Nat. Acad. Sci., U.S.A.* **52**, 1009–1017.

Sjöberg, B. (1977). *Eur. J. Biochem.* **81**, 277–283.

Smillie, L. B. & Hartley, B. G. (1966). *Biochem. J.* **101**, 232–241.

Summers, M. R., Smythers, G. W. & Oroszlan, S. (1973). *Anal. Biochem.* **53**, 624–628.

Unge, T. & Strandberg, B. (1979). *Virology*, **96**, 80–87.

Unge, T., Liljas, L., Strandberg, B., Vaara, I., Kannan, K. K., Fridborg, K., Nordman, C. E. & Lentz, P. J. (1980). *Nature (London)*, **285**, 373–377.

van Emmelo, J., Devos, R., Ysebaert, M. & Fiers, W. (1980). *J. Mol. Biol.* **143**, 259–271.

Weber, K. & Osborn, M. (1975). *The Proteins* (Neurath, H., Hill, R. L. & Boeder, C. L., eds), 3rd edit., vol. 1, pp. 179–223, Academic Press, New York, San Francisco, London.

Ysebaert, M., van Emmelo, J. & Fiers, W. (1980). *J. Mol. Biol.* **143**, 273–287.

Edited by S. M. Weissman