PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Studies on the Structure of the Haemagglutinin

J. J. Skehel, M. D. Waterfield, J. W. McCauley, K. Elder and D. C. Wiley

Phil. Trans. R. Soc. Lond. B 1980 **288**, 335-339
doi: 10.1098/rstb.1980.0009

Phil. Trans. R. Soc. Lond. B **288**, 335–339 (1980)
Printed in Great Britain

Studies on the structure of the haemagglutinin

By J. J. Skehel,† M. D. Waterfield,‡ J. W. McCauley,†
K. Elder‡ and D. C. Wiley§
† *Division of Virology, National Institute for Medical Research, Mill Hill, London NW7 1AA, U.K.*
‡ *Imperial Cancer Research Fund, Lincoln's Inn Fields, London W.C.2, U.K.*
§ *Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts* 02138, *U.S.A.*

The biosynthesis of the haemagglutinin glycoproteins of infectious influenza virus particles involves proteolytic cleavage of the primary translation products and the amino acid sequences at the two sites of processing are presented. In addition, details of the primary structure of the haemagglutinin of A/Japan/305/57 (H2N1) are reported and compared with information available for haemagglutinins of other subtypes.

The influenza virus haemagglutinin appears to be a trimer of molecular mass 210000 in which each monomer consists of two disulphide linked glycosylated polypeptides HA_1 and HA_2 (see Waterfield *et al.* 1979).

During virus replication, HA_1 and HA_2 are formed as the result of proteolytic cleavage of a precursor glycopolypeptide of molecular mass about 70000, a process which has been shown to be essential for the formation of infectious virus particles even though in its absence, particles that which can bind to cells can be assembled (see Rott *et al.*, this symposium). The order of biosynthesis of HA_1 and HA_2 in the precursor glycopolypeptide – NH_2–HA_1–HA_2–COOH – was deduced from comparative amino terminal sequence analyses of the three glycopolypeptides in which it was found that the amino terminal sequence of HA_1 and the precursor were identical. The amino terminus of HA_2 is, therefore, generated by the cleavage process (Skehel & Waterfield, 1975). Analyses of the translation of haemagglutinin messenger RNA *in vitro*, however, indicate that this is not the only proteolytic cleavage of the primary translation product that occurs during biosynthesis. In fact, like secreted proteins (Blobel & Dobberstein 1975; Rothman & Lodish 1977) and the membrane glycoprotein of vesicular stomatitis virus (Lingappa *et al.* 1978), the nascent haemagglutinin contains a signal peptide at its amino terminus which may be involved in the initial association with and the eventual transfer of the protein through the lipid bilayer of the endoplasmic reticulum. As a consequence, the amino terminus of HA_1 is also generated by a process involving proteolysis since the signal peptide is removed from the completed glycopolypeptide chain. Evidence for this process and for the existence of the signal peptide was obtained from experiments on the translation of haemagglutinin messenger RNA in wheat germ cell-free protein synthesizing systems in which unprocessed haemagglutinin polypeptides accumulate (Elder *et al.* 1979). The amino terminal sequence of the signal-containing polypeptide was obtained by analysing the polypeptides produced in the presence of defined radioactive amino acid precursors and is shown in table 1. From these data it appears that like other signal peptides that of the haemagglutinin is rich

in hydrophobic amino acids and also that since isoleucine residues of equivalent spacing are present at positions 3 and 5 of HA_1 and 17 and 19 of the in-vitro product the signal peptide appears to be 14 amino acids long.

TABLE 1. THE AMINO TERMINAL SEQUENCES OF A/JAPAN/305/57 (H2) HAEMAGGLUTININ SYNTHESIZED *IN VITRO*

	−14	−10	−5	1	5
unprocessed *in vitro*	? -Ile-Ile- ?	-Leu-Ile-Leu-Leu- ?	-Ile- ?	-Val- ? - ? - ?	- ? -Ile- ? -Ile- ? -
processed *in vitro*				? - ? -Ile- ?	-Ile- ? -
HA_1 from infectious virus				Asp-Gln-Ile-Cys-Ile-Gly-	
HA_2 from infectious virus				Gly-Leu-Phe-Gly-Ala-Ile-	

As stated above, the haemagglutinin molecule appears from the results of crosslinking experiments to be a trimer of identical subunits (Wiley *et al.* 1977). This conclusion is consistent with the molecular mass estimates of 210000 for the protease-released molecule and 70000 for each monomer, which were obtained by ultracentrifugation and amino acid sequence analyses.

Evidence for the orientation of the haemagglutinin polypeptides with respect to the bilayer of the virus particle comes from comparative molecular mass, compositional and amino acid sequence analyses of detergent and protease released haemagglutinins (Skehel & Waterfield 1975) and also from labelling experiments with photoactivatable hydrophobic azides (Wiley & Skehel 1978). The results of both types of experiment indicate that the site of lipid association is near the carboxyl-terminus of HA_2.

Two other groups have been involved in determining the amino acid sequences of various haemagglutinins; these groups are D. Bucher and her coworkers at Mount Sinai, New York (Bucher *et al.* 1976), and C. Ward and T. Dopheide (Dopheide & Ward 1978; Ward & Dopheide 1979) at the C.S.I.R.O., Melbourne, and the rest of this communication involves a description of the data that we have obtained and a comparison of them with those available from the other groups.

The haemagglutinin on which we have concentrated is of the H2 subtype isolated from the A/Japan/305 strain of 1957, this was chosen primarily because the amino terminus of the HA_1 component of this molecule is not blocked and the spacing of the methionine residues allows the production of peptides suitable for automatic sequence determination after cleavage with cyanogen bromide. Selected regions of the sequence are shown in tables 2, 3, 4 and 5. HA_1 is made up of about 316 amino acids and HA_2 of about 210. These figures cannot be established with certainty until the sequence is completed. HA_1 contains six methionine residues and HA_2 contains thirteen, seven of which may be associated with the hydrophobic lipid associated region. Both of the polypeptides are glycosylated and there are sites of glycosylation at residues 11, 23, 163 and 277 in HA_1 and a single site at residue 156 in HA_2. In each case the carbohydrate side chains contain *N*-acetylglucosamine, galactose, mannose and fucose and are, therefore, of the complex type and in each case also they are attached by *N*-acetylglucosamine–asparagine linkages in the characteristic sequence of asparagine–X–serine or threonine (Marshall & Neuberger 1970). This appears to be so for all the carbohydrate side chains in all haemagglutinins examined although in some haemagglutinins, side chains of the simple type containing only *N*-acetylglucosamine and mannose have been detected

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

(Ward & Dopheide 1979). The number of cysteine residues in the haemagglutinin of A/Japan/305/57 has been difficult to establish but it is likely that HA_1 contains nine and HA_2 five residues. At present we do not know the location of the disulphide linkages, although the two cysteine residues at 265 and 269 are of interest since two of the cysteine residues in HA_2 have a similar spacing and this may be indicative of interchain linkage.

TABLE 2. CARBOXYL TERMINAL SEQUENCES OF HA_1 GLYCOPOLYPEPTIDE CHAINS

A/Bel/42 (H0)	(Asx, Ile, Pro, Ser, Ser, Glx, Ile)
A/Weiss/43 (H0)	(Asx, Ile, Pro, Ser, Ser, Glx, Ile)
A/FM/1/47 (H1)	(Asx, Ile, Pro, Ser, Ser, Glx, Ile)
A/Japan/305/57 (H2)	Asx-Val-Pro-Glx-Ser-Glx-Ile
X-31 (H3)	(Asx, Val, Pro, Glx, Lys) (Glx, Thr)
A/Memphis/72 (H3)†	Asn-Val-Pro-Glu-Lys-Glu-Thr
MRC-11 (H3)	(Asx, Val, Pro, Glx, Lys) (Glx, Thr)

† From Dopheide & Ward (1978).

The carboxyl terminus of HA_1 is in all cases either isoleucine or threonine (table 2). Since the infectivity of particles containing uncleaved precursor haemagglutinins can be activated by trypsin it can be deduced that a specific carboxyl terminus is not essential for infectivity and also that there is at least one basic residue between the HA_1 and HA_2 polypeptide chains. It also appears possible from these results that an exopeptidase may be involved in carboxyl terminal processing *in vivo*.

Comparative sequence analyses allow the conclusion that certain features of both glycopolypeptides are conserved from strain to strain of virus. In addition to the common lack of a basic residue at the carboxyl termini of the HA_1 components, there is considerable conservation of sequence in this region of the molecule (table 2) which in fact extends for about 25 residues (Waterfield *et al.* 1979).

Similar data showing small regions of conservation are also available for the amino-terminal regions of HA_1 and are shown in table 3.

The most extensive comparison that can be made at the moment, however, is between the sequences of HA_1 components from viruses of the H2 and H3 subtypes; these are shown in table 4. The sequence involves 80 residues at the carboxyl terminus of which only 38 are identical; 13 of the substitutions involve more than one base change. They are clearly quite different, a situation which contrasts sharply with that observed if the sequences of two H3 molecules are compared. For example, if the 1968 Hong Kong haemagglutinin is compared with that of this 1972 strain in this region from residue 257 to 316, only 5 out of the 59 residues vary (Waterfield *et al.* 1979).

However, even between haemagglutinins of the different subtypes there are also conserved features most clearly at the carboxyl terminus of HA_1 as mentioned above (table 2) and in the region of the cysteine residues that have exactly the same locations. Interestingly, also, the site of glycosylation is at position 273 in HA_1 of the H3 subtype as opposed to position 277 for HA_1 of the H2 subtype.

Finally, the site of sequence similarity first observed and the most extensive region of conservation among all the haemagglutinins of different subtypes examined is at the amino-terminus of HA_2 (table 5). Of the first 20 residues, 17 are identical. From a structural point of view the presence of glycines at residues 1, 4, 8, 12, 13, 16, 20, 23 and 31 is of interest and

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

TABLE 3. AMINO TERMINAL SEQUENCES OF HA_1 GLYCOPOLYPEPTIDE CHAINS

A/Bellamy/42 (H0)	Asx-Thr-Ile-Cys-Ile-Gly-Tyr-His-Ala-Asx
A/Weiss/43 (H0)	Asx-Thr-Ile-Cys-Ile-Gly-Tyr-His-Ala-Asx
	CHO CHO
A/Japan/305/57 (H2)	Asp-Gln-Ile-Cys-Ile-Gly-Tyr-His-Ala-Asx-Asn-Thr-Ser-Glu-Lys-Val-Asp-Thr-Ile-Leu-Glu-Arg-Asn-Val-Thr-Thr-His
X-31 (H3)	blocked
MRC-11 (H3)	blocked
X-38 (Heq1)†	Asp-Lys-Ile-Ser-Leu-Gly-Tyr-His-Ala-Val
B/Lee/40	Asx-Thr-Ile

† From Bucher *et al.* (1976).

TABLE 4. AMINO ACID SEQUENCES OF THE CARBOXYL TERMINAL REGIONS OF HA_1 GLYCOPOLYPEPTIDE CHAINS

A/Japan/305/57 (H2)	Gly-Asn-Leu-Ile-Ala-Pro-?-Glu-Tyr-Phe-Lys-Ile-Ser-Lys-Arg-Gly-Ser-Ser-Gly-Ile-Met-Lys-Thr-Glu-Gly-Thr-Leu-Glu-Asn
A/Memphis/72 (H3)	Gly-Asx-Leu-Ile-Ala-Pro-Arg-Gly-Tyr-Phe-Lys-Met-Arg-Thr-Gly-Lys-Ser-Ser – Ile-Met-Arg-Ser-Asp-Ala-Pro-Ile-Gly-Thr
	CHO
A/Japan/305/57	Cys-Glu-Thr-Lys-Cys-Glu-Thr-Pro-Leu-Gly-Ala-Ile-Asn-Thr-Thr-Leu-Pro-Phe-His-Asx-Val-His-Pro-Leu-Thr-Ile-Gly-Glx-Cys
	CHO
A/Memphis/72	Cys-Ile-Ser-Glu-Cys-Ile-Thr-Pro-Asn-Gly-Ser-Ile-Pro-Lys-Pro-Asp-Asp-Phe-Gln-Asn-Val-Asn-Lys-Ile-Thr-Tyr-Gly-Ala-Cys
A/Japan/305/57	Pro-Lys-Tyr-Ser-Glu-Lys-Leu-Val – Leu-Ala-Thr-Gly-Leu-Arg-Asx-Val-Pro-Glx-Ser-Glx-Ile
A/Memphis/72	Pro-Lys-Tyr-Val-Lys-Gln-Asn-Thr-Leu-Lys-Leu-Ala-Thr-Gly-Met-Arg-Asn-Val-Pro-Glu-Lys-Gln-Thr

TABLE 5. AMINO ACID SEQUENCES OF THE AMINO TERMINAL REGIONS OF HA_2 GLYCOPOLYPEPTIDE CHAINS

	1 5 10 15
A/Bellamy/42 (H0)	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Gly-Gly-Trp-Thr-Gly-Met
A/Weiss/43 (H0)	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Gly-Gly-Trp-Thr-Gly-Met
	20 25 30
A/Japan/305/57 (H2)	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Gly-Gly-Trp-Glu-Gly-Met-Val-Asp-Gly-Trp-Tyr-Gly-Tyr-?-Tyr-His-Asx-Asx-Glu-Gly
A/Korea/67 (H2)	(Gly, Leu, Phe, Gly, Ala, Ile, Ala, Gly, Phe, Ile, Glu, Gly, Gly, Trp, Glu, Gly, Met, Val, Asp, Gly, Trp, Tyr, Gly, Tyr, Ser, Tyr, His, Asx, Asx, Gly, Gly)
A/X-31/68 (H3)	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Met-Ile-Asp-Gly-Trp-Tyr-Gly-Phe-Arg-His-Gln-Asn-Ser-Glu-Gly
A/Memphis/72 (H3)†	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Met-Ile-Asp-Gly-Trp-Tyr-Gly-Phe-Arg-His-Gln-Asn-Ser-Glu-Gly
A/M.R.C.11/73 (H3)	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Met
X-38 (Heq 1)‡	Gly-Leu-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Ile-Glu-Asn-Gly-Trp-Glu-Gly-Leu
B/Lee/40	Gly-Phe-Phe-Gly-Ala-Ile-Ala-Gly-Phe-Leu-Glu-Gly-Gly-Trp-Glu-Gly-Met

† From Ward & Dopheide (1979). ‡ From Bucher *et al.* (1976).

also since all haemagglutinins appear to have characteristic circular dichroism spectra in the near ultraviolet, which is not detected for the uncleaved molecules containing inactive precursor (Flanagan & Skehel 1977), it is of some interest that there are a number of aromatic residues in this region near the site of proteolytic cleavage at positions 3, 9, 14, 21, 22, 24 and 26. It was also initially striking that residues 3–9 were in a pseudopalindrome about the isoleucine at position 6 but it may be more significant to function that all ten of these residues are uncharged and hydrophobic and are in an analogous sequence to the amino terminus of one of the polypeptides of Sendai virus fusion glycoprotein which is also generated during activation of fusion activity by proteolysis (Gething *et al.* 1978).

Apart from this region of the HA_2 chain and around the cysteine residues at positions 144 and 148 there is very little conservation and this would certainly seem to rule out the possibility of constant and variable chain components for the haemagglutinin.

In summary, then, the overall picture that is obtained from these sequence studies is that haemagglutinins of the same subtype are extremely similar and this applies to both HA_1 and HA_2 polypeptide components. There are, however, large differences again in both polypeptide chains between the haemagglutinins of different subtypes and these data would therefore not seem to support those theories on the mechanism of antigenic shift which imply that each new subtype is an extension of the previous subtype.

References (Skehel *et al.*)

Blobel, G. & Dobberstein, B. 1975 *J. Cell Biol.* **67**, 835–851.

Bucher, D. J., Li, S. S. L., Kehoe, J. M. & Kilbourne, E. D. 1976 *Proc. natn. Acad. Sci. U.S.A.* **73**, 238–242.

Dopheide, T. A. A. & Ward, C. W. 1978 In *The influenza virus haemagglutinin* (ed. W. G. Laver, H. Bachmayer & R. Weil) (*Topics in infectious diseases*, vol. 3), pp. 193–201. Vienna: Springer-Verlag.

Elder, K. T., Bye, J. M., Waterfield, M. D., Skehel, J. J. & Smith, A. E. 1979 *Virology* **95**, 343–350.

Flanagan, M. T. & Skehel, J. J. 1977 *FEBS Lett.* **80**, 57–60.

Gething, M. J., White, J. M. & Waterfield, M. D. 1978 *Proc. natn. Acad. Sci. U.S.A.* **75**, 2737–2740.

Lingappa, V. R., Katz, F. N., Lodish, H. F. & Blobel, G. 1978 *J. biol. Chem.* **253**, 8667–8670.

Marshall, R. D. & Neuberger, A. 1970 *Adv. Carbohyd. Chem. Biochem.* **25**, 407–478.

Rothman, J. E. & Lodish, H. F. 1977 *Nature, Lond.* **269**, 775–780.

Skehel, J. J. & Waterfield, M. D. 1975 *Proc. natn. Acad. Sci. U.S.A.* **72**, 93–97.

Ward, C. W. & Dopheide, T. A. A. 1979 *Br. med. Bull.* **35**, 51–56.

Waterfield, M. D., Espelie, K., Elder, K. & Skehel, J. J. 1979 *Br. med. Bull.* **35**, 57–63.

Wiley, D. C., Skehel, J. J. & Waterfield, M. D. 1977 *Virology* **79**, 446–448.

Wiley, D. C. & Skehel, J. J. 1978 In *The influenza virus haemagglutinin* (ed. W. G. Laver, H. Bachmayer & R. Weil) (*Topics in infectious diseases*, vol. 3), pp. 135–138. Vienna: Springer-Verlag.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES