Nucleotide Sequence Analysis of the Haemagglutinin–Neuraminidase Gene of Newcastle Disease Virus

By NEIL S. MILLAR, PHILIP CHAMBERS AND PETER T. EMMERSON*
Department of Biochemistry, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, U.K.

(Accepted 27 May 1986)

SUMMARY

The nucleotide sequence of the haemagglutinin–neuraminidase (HN) gene of Newcastle disease virus (NDV) has been determined. The HN gene is 2031 nucleotides long, approximately 13.5% of the viral genome. The nucleotide sequence contains a single long open reading frame which would encode a protein of 577 amino acids, with a mol. wt. of 63149. This is in good agreement with estimates of the molecular weight of the unglycosylated HN protein. Analysis of the amino acid sequence reveals six potential glycosylation sites and shows the major hydrophobic region to be close to the N terminus. This provides evidence for the N-terminal attachment of HN to the viral membrane. The hydrophilic nature of the extreme N-terminal amino acids suggests the absence of a cleaved signal sequence. Analysis of the long non-coding region at the 3' end of the mRNA encoded by the HN gene of NDV suggests a possible explanation for the origin of HN₀ in extremely avirulent strains of NDV. There are regions of high homology between the deduced amino acid sequence of the NDV HN glycoprotein and the HN glycoproteins of two other paramyxoviruses, Sendai virus and simian virus 5 (SV5). An alignment of the HN amino acid sequences of these viruses shows 32% of amino acid residues are conserved between NDV and SV5, and 23% between NDV and Sendai virus. In contrast, only very limited homology is found between NDV HN and the influenza virus glycoproteins.

INTRODUCTION

Newcastle disease virus (NDV) is the prototype virus of the Paramyxovirus genus. NDV infects poultry and a variety of other avian species causing fowl pest, a disease of considerable agricultural and economic significance. The severity of infection is, however, highly strain-dependent, but in severe cases results in fatal haemorrhagic intestinal lesions or acute respiratory and nervous problems (Box, 1967; Alexander, 1985). NDV is an enveloped virus and contains two surface glycoproteins, the haemagglutinin–neuraminidase (HN) and fusion (F) proteins. The HN glycoprotein contains both the haemagglutinating and neuraminidase activities of the virus which are responsible for initial attachment of the virus particle to its cellular receptor and receptor-destroying activity, respectively (Scheid & Choppin, 1973). The F glycoprotein is involved in virus penetration, cell fusion and haemolysis (Homma & Ohuchi, 1973; Nagai et al., 1976). The virus also contains four internal structural proteins. These are the matrix (M) protein located on the inside of the viral envelope and the nucleocapsid protein (NP), phosphoprotein (P) and large protein (L) which form the transcriptive complex (Hamaguchi et al., 1983), which together with the viral genomic RNA (a negative sense, single-stranded RNA of approximately 15000 bases), forms the virion nucleocapsid. The NDV genome also codes for two non-structural proteins, 36K and 33K, probably on the same mRNA as P (Chambers & Samson, 1982; Collins et al., 1982). Molecular cloning of cDNA to 90% of the NDV genome (Chambers et al., 1986) has established the gene order as 3'–NP–P–M–F–HN–L–5'.
The infectivity of NDV is neutralized by either monoclonal antibodies or monospecific antisera to the HN glycoprotein (Iorio & Bratt, 1984; Umino et al., 1984). Monoclonal antibodies have been used to show that the haemagglutinin and neuraminidase sites of HN are antigenically separate (Iorio & Bratt, 1984). The nucleotide sequences of the HN and F genes are of particular interest since the pathogenicity of NDV appears to be largely dependent on the structure of the two surface glycoproteins and their susceptibility to proteolytic cleavage (Nagai et al., 1976). Activation of the paramyxovirus F glycoprotein requires the post-translational cleavage of a precursor F₀ into two disulphide-linked fragments F₁ and F₂ (Scheid & Choppin, 1977). In avirulent strains of NDV, such as La Sota and B₁, F₀ is not cleaved in certain cell lines resulting in decreased viral infectivity (Nagai et al., 1976; Garten et al., 1980a). The paramyxovirus HN glycoprotein is not thought to require modification by post-translational cleavage except in the case of two extremely avirulent strains of NDV (Ulster and Queensland). These strains, like La Sota and B₁, encode F₀ which is not cleaved in certain cell lines and also encode a larger precursor (HN₀) of the HN molecule (Nagai et al., 1976; Nagai & Klenk, 1977). In strains Ulster and Queensland the proteolytic cleavage of a C-terminal glycopeptide from HN₀ is necessary for the generation and activation of HN (Garten et al., 1980b).

The evidence currently available suggests that the paramyxovirus F glycoprotein is attached to the membrane near its C-terminal end and has a hydrophobic signal sequence cleaved from its N terminus (Paterson et al., 1984; Blumberg et al., 1985 a), as is common for viral glycoproteins (von Heijne, 1983; Wiley, 1985). The HN glycoprotein, however, seems to be attached near its N-terminal end and does not have a cleaved signal sequence (Schuy et al., 1984; Blumberg et al., 1985b; Hiebert et al., 1985a). In this respect, it is similar to the respiratory syncytial virus G protein (Satake et al., 1985; Wertz et al., 1985) and influenza virus neuraminidase (Fields et al., 1981; Blok et al., 1982).

METHODS

Molecular cloning of HN. Details of the molecular cloning of cDNA: RNA hybrids to NDV virion RNA and of the methods used to map the clones to their position in the viral genome, namely dot blot hybridization, Northern blot hybridization, colony hybridization and restriction enzyme mapping, have been published previously (Chambers et al., 1986). Five clones designated 1.13, 2.73, 3.73, 4.68 and 7.44 (constructed as cDNA inserted into the PstI site of pBR322) cover the entire HN gene.

DNA sequencing. The cDNA inserts from the five clones identified as covering the HN gene were subcloned into the phage vectors M13 mp8, mp9, mpl8 or mpl9 (Messing & Vieira, 1982; Yanisch-Perron et al., 1985) using a variety of restriction enzyme sites to 'force-clone' fragments of the cDNA inserts. DNA sequencing was performed by the dideoxy chain-termination method (Sanger et al., 1977) using a universal 17 base primer (Pharmacia) and [α-32P]dATP as label. Thin polyacrylamide buffer gradient gels were used as described by Biggin et al. (1983) and exposed to Fuji RX X-ray film. The sequencing data were stored and assembled to give a consensus sequence by use of the DB system of computer programs (Staden, 1982b, 1986). Analysis of the sequence was aided by the use of the ANALYSEP, ANALYSEQ, DIAGON and Microgenie programs (Staden, 1982a, 1984, 1986; Queen & Korn, 1984).

RESULTS

Nucleotide sequence of the HN gene

The relative positions of cDNA clones covering the HN gene and the location of restriction sites used to 'force-clone' cDNA fragments into M13 vectors are shown in Fig. 1. Restriction sites were chosen to produce overlapping sequencing data and to allow the sequence to be determined on both strands of the cDNA.

The cDNA sequence, in the positive (mRNA) sense, of a 2265 nucleotide region is presented in Fig. 2. The region shown extends from an SphI site 5' to the deduced HN mRNA start to an RsaI site 3' to its deduced end. The probable position of the HN and L mRNA start sites and the polyadenylation signal sequence at the ends of the F and HN mRNAs are indicated. These regions were located largely on the basis of comparisons with the mRNA start and polyadenylation sites of related viruses (Rose, 1980; Gupta & Kingsbury, 1984), as shown in Fig. 3 and in a previous report (Chambers et al., 1986), and also on provisional S1 mapping experiments (data not shown). There appears to be an intergenic trinucleotide between the HN
and L genes (similar to Sendai virus) and a dinucleotide between F and HN (as occurs in the less closely related vesicular stomatitis virus). The HN mRNA would, therefore, have non-coding regions of 119 and 180 nucleotides at its 5' and 3' ends, respectively. The possible significance of the relatively long non-coding regions is discussed later.

**Analysis of the HN glycoprotein amino acid sequence**

The HN gene nucleotide sequence contains only one long open reading frame. This starts at the ATG codon at nucleotides 306 to 308 and continues to a termination codon at nucleotides 2037 to 2039 which is followed by a second termination codon, in phase, at nucleotides 2052 to 2054. The start of translation would be difficult to determine directly by amino acid sequencing due to the blocked N terminus of the HN glycoprotein (Scheid et al., 1978; Schuy et al., 1984). However, the ATG codon at the start of the long open reading frame is flanked by nucleotides consistent with those identified by Kozak (1981, 1986) as characteristic of eukaryotic protein initiation sequences (5'-purine-N-N-A-U-G-G-3'). This region would encode a protein of 577 amino acids with a predicted mol. wt. of 63149. This in good agreement with estimates of the mol. wt. of the unglycosylated form of HN (67K) when synthesized in a cell-free system (Clinkscales et al., 1977), or in NDV-infected cells in the presence of tunicamycin (Morrison & Simpson, 1980). The mol. wt. of the glycosylated protein has been estimated to be 74K on SDS–polyacrylamide gels, which suggests the presence of only three or four oligosaccharides per HN monomer, since these usually add 2000 to 3000 mol. wt. per oligosaccharide chain to the mol. wt. of a protein (Keil et al., 1979; Horisberger et al., 1980). Six potential asparagine-linked glycosylation sites are present in the protein and are shown as boxed regions in the amino acid sequence (Fig. 2). Two of the six potential sites may be less favourable due to the presence of proline or aspartic acid at position X in the sequence Asn–X–Ser/Thr (Ronin et al., 1978; Hart et al., 1979). However, of the six sites, Asn–Pro–Thr (amino acids 500 to 502) is the only one which appears to be conserved between NDV, SV5 and Sendai virus (see Fig. 6). Most of the potential carbohydrate attachment sites of the NDV HN are in the C-terminal half of the molecule, as is the case for Sendai virus and SV5 (Shioda et al., 1986).

A hydropathy plot of the predicted amino acid sequence of HN (Fig. 4b) shows that the major hydrophobic region of the protein is located towards the N terminus (amino acids 27 to 54) and that the extreme N-terminal amino acids (1 to 26) are relatively hydrophilic. No long stretches of hydrophobic amino acids can be detected near its C terminus. These features suggest that the HN glycoprotein is attached in the viral membrane at a site close to its N rather than C terminus and indicate the absence of a cleaved signal sequence since the N-terminal region does not show
**Fig. 2.** Nucleotide sequence and deduced amino acid sequence of the NDV HN gene. The cDNA sequence is shown in the positive (mRNA) sense, and covers a 2265 base region of the NDV genome. The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown in the positive (mRNA) sense, and covers a 2265 base region of the NDV genome. The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown as boxed regions in the deduced amino acid sequence.

The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown as boxed regions in the deduced amino acid sequence.

The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown as boxed regions in the deduced amino acid sequence.

The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown as boxed regions in the deduced amino acid sequence.

The sequence is numbered from the reading frame is shown above the nucleotide sequence. Potential asparagine-linked glycosylation sites are shown as boxed regions in the deduced amino acid sequence.
Fig. 3. Nucleotide sequences of the NDV F/HN and HN/L gene junctions are shown together with the start of the NP mRNA determined by Kurilla et al. (1985) and the M/F gene junction (unpublished data from this laboratory). A consensus sequence is shown based on these data. The NDV sequences are also compared to the consensus mRNA start and termination sequences of Sendai virus and vesicular stomatitis virus (VSV).

Fig. 4. Charge and hydropathy profiles of the HN sequence. (a) A window of 11 amino acids was used to calculate charge distribution. Basic regions of HN are indicated by peaks above, and acidic regions by peaks below the midline. (b) A window of 11 amino acids was used to calculate the hydropathy profile, using the procedure of Kyte & Doolittle (1982). Hydrophobic regions of HN are indicated by peaks above, and hydrophilic regions by peaks below the horizontal line which represents the average hydropathy of sequenced proteins. The HN amino sequence is numbered from its N to C terminus on the horizontal scale.
the features identified by von Heijne (1983) as being necessary for cleavage. Fig. 4(a) shows the distribution of charged amino acids in HN. NDV HN has an approximately neutral isoelectric point (Chambers & Samson, 1982) but there is an asymmetric distribution of acidic and basic residues within the protein, the N-terminal half being acidic and the C-terminal half basic. The C terminus of HN is therefore the probable origin of the highly basic 40K polypeptide, gp40, a cleavage product of HN observed in cells infected with certain strains of NDV in vivo (Chambers & Samson, 1982).

Comparison of the HN sequences of NDV, SV5 and Sendai virus

The amino acid sequences of NDV, SV5 and Sendai virus HN glycoproteins were compared by using the dot matrix homology computer program DIAGON (Staden, 1984). The plots show considerable similarity between these sequences (Fig. 5). Shioda et al. (1986) have identified two regions of 20 amino acids as being particularly well conserved between the Sendai virus and SV5 HN glycoproteins. By comparing the NDV HN amino acid sequence with those of SV5 (Hiebert et al., 1985a) and two strains of Sendai virus (Blumberg et al., 1985b; Miura et al., 1985; Shioda et al., 1986) we have been able to produce an alignment showing homology over much of the HN glycoprotein (Fig. 6). Insertions and deletions of amino acids are required to permit the alignment of regions of high homology. In the proposed alignment, all 12 of the cysteine residues in NDV HN correspond with cysteines in the SV5 sequence but only ten with Sendai virus. Regions of the HN sequence are particularly well conserved between these viruses. For example, a 100 amino acid region of NDV HN (residues 167 to 267) shows 54% identity with SV5 and 39% with Sendai virus. Overall, 32% of amino acids are conserved between NDV and SV5 and 23% between NDV and Sendai virus. Shioda et al. (1986) reported 24% conservation of amino acids between SV5 and Sendai HN glycoproteins, similar to the 22% we obtain. Preliminary results obtained with the NDV F glycoprotein show very similar values for overall homology with SV5 and Sendai virus.

DISCUSSION

The HN gene product of NDV strain Beaudette C is a protein of 577 amino acids with a mol. wt. of 63K. There are six potential asparagine-linked glycosylation sites, two of which may, however, be less favourable as discussed above. All but one of the potential glycosylation sites are in the C-terminal half of the molecule, a feature which Shioda et al. (1986)
NDV HN gene sequence

**Fig. 6.** Comparison of the HN amino acid sequences of NDV, SV5, and Sendai virus. The HN sequences have been aligned to show SV5 above and Sendai virus below the predicted NDV sequence. The sequence of Sendai virus strain Harris (S.H) is shown, excluding the first 34 amino acids where no significant homology to NDV or SV5 was detected. Below this is the sequence of strain Z (S.Z), shown only where it differs from strain Harris. The HN sequence of strain Z differs from strain Harris, at positions 148 (aspartic acid) and 238 (isoleucine), in only one of the two published versions (Miura et al., 1985; Shioda et al., 1986). At position 533, the deletion of one amino acid in strain Z with respect to strain Harris is indicated by a dash. The amino acid sequences are numbered from their N to C termini, shown on the right-hand side of the figure.

mention as being common to Sendai virus and SV5. HN of NDV strain Beaudette C is cleaved *in vivo* to produce a glycosylated 40K polypeptide, designated gp40, which has been shown by fluorography to contain a large proportion of the HN carbohydrate (Chambers & Samson, 1982). This cleavage product has a highly basic isoelectric point and it was, therefore, assumed that the other region of HN would be acidic since intact HN has a neutral isoelectric point. Our results (Fig. 4a) demonstrate that the C-terminal half of NDV HN is basic and contains all but one of the glycosylation sites and that the N-terminal half is acidic. This suggests that gp40 is derived from the C-terminal half of HN. The stability of gp40 *in vivo* suggests that it is a protease-resistant domain of the HN polypeptide.
Evidence to suggest a proposed N- rather than C-terminal attachment of the NDV HN glycoprotein in the viral membrane was provided by Schuy et al. (1984). A 9K fragment removed from HN₀ of intact NDV virions (strain Ulster) was shown by N- and C-terminal amino acid analysis to be derived from the C terminus of the molecule, while the N termini of both HN₀ and HN remained blocked. A hydropathy plot of NDV HN (Fig. 4b) supports this interpretation by showing that the major hydrophobic region of the protein is located near the N terminus of HN. The relatively hydrophilic nature of the extreme N-terminal amino acids suggests the absence of a signal sequence which is cleaved after translation (von Heijne, 1983). This feature appears to be common among N-terminally attached viral glycoproteins (Fields et al., 1981; Blok et al., 1982; Blumberg et al., 1985b; Hiebert et al., 1985a; Satake et al., 1985). The hydrophilic region at the extreme N terminus of HN probably represents the cytoplasmic tail of the molecule.

A feature of the NDV HN gene sequence discussed in a previous publication (Chambers et al., 1986) is the presence of a relatively long non-coding region at the 3' end of the mRNA which has the potential to code for up to 55 extra amino acids. The nucleotide sequence reveals the presence of a further potential glycosylation site in this region in phase with the open reading frame (nucleotides 2103 to 2111). This may help to explain the existence of the precursor HN₀ in the strains Ulster and Queensland, since HN₀ is cleaved to active HN by the removal of a C-terminal glycopeptide (Garten et al., 1980b). The cloning and sequencing of strain Ulster is now underway in this laboratory and it is hoped that this will provide further information about HN₀.

Hiebert et al. (1985b) reported a previously unrecognized gene in SV5 encoding a small hydrophobic (SH) protein of 44 amino acids, located between the F and HN mRNAs. There is a small open reading frame of 41 amino acids near the proposed start of the NDV HN gene (nucleotides 197 to 319) but if translated this would give a basic protein which is considerably more hydrophilic than SH of SV5. Shortly after the end of this reading frame, which overlaps the start of the HN coding region, is a sequence (nucleotides 356 to 374) which shows some similarity to the NDV mRNA conserved start and polyadenylation sequences (Fig. 3). Due to its small size and highly basic nature, such a protein would not have been detected on the two-dimensional gels normally used for the analysis of NDV proteins (Chambers & Samson, 1982). At present, there is no evidence for the generation of such a protein.

Comparisons of the NDV HN sequence with those of Sendai virus and SV5, aided by the use of dot matrix homology plots (Fig. 5), enabled the identification of conserved regions between the viruses (Fig. 6). However, regions of high sequence homology between NDV HN and the influenza virus glycoproteins have not been detected by computer analysis or by eye. Blumberg et al. (1985b) compared the Sendai virus HN sequence with the influenza virus HA and NA sequences. A 200 amino acid region of Sendai virus HN, located in the centre of its primary sequence, was aligned with the influenza A and B viruses' neuraminidases (NA), and a separate 100 amino acid region towards the C-terminal end of HN was aligned with influenza A, B and C viruses' haemagglutinins (HA). This alignment was used as evidence for the possible location of the two active sites within the HN molecule. However, in the case of the HA sequences, only the alignment to influenza C virus could be considered significant and the alignment with the NA sequences were statistically indecisive. We have found that few of the residues identified as being conserved between Sendai virus HN and influenza virus HA or NA appear to be conserved among the three paramyxoviruses. In the neuraminidase region only 16 out of 60, and in the haemagglutinin region 12 out of 30 residues are common to two or more of the paramyxoviruses. We have identified a conserved region between NDV, SV5 and Sendai virus [Gly–Ala–Glu–Gly–Arg–(Leu/Ile)] at positions 399 to 404 in the NDV HN amino acid sequence which shows similarity to the influenza haemagglutinin sialic acid receptor-binding site (Rogers et al., 1983). The homology is greatest between NDV (an avian paramyxovirus) and avian influenza viruses, which have the sequence Gly–Gln–Ser–Gly–Arg–Ile (Naeve et al., 1984).

A secondary structure prediction of the three paramyxovirus HN sequences using the method of Garnier et al. (1978) suggested that the C-terminal half of the sequences almost exclusively favour β-sheet and reverse turn structures (data not shown). This is the region of the NDV HN protein which corresponds to the stable gp40 fragment. It may therefore be a folded domain.
analogue to either the HA1 region of influenza virus HA or to the influenza virus NA head region, both of which have been shown by X-ray crystallography to be globular regions of antiparallel β-sheet (Wilson et al., 1981; Varghese et al., 1983). The high level of predicted β-sheet structure in the central region of Sendai virus HN, identified by Blumberg et al. (1985b) and compared to influenza virus NA, is not so apparent in NDV HN. The similarity of the conserved six amino acid paramyxovirus sequence to the receptor-binding site of influenza virus HA (mentioned above) may, however, support the localization of the haemagglutinin site to the C-terminal region of HN, as suggested by Blumberg et al. (1985b).

Viral gene sequences may provide information about evolutionary relationships between NDV and members of the Paramyxoviridae and other viral families. Details of this kind are now emerging between positive strand RNA viruses (Haseloff et al., 1984). Alignment of the NDV, SV5 and Sendai virus HN sequences (Fig. 6) suggests that NDV is more closely related to SV5 than to Sendai virus (overall amino acid conservation of 32% and 23%, respectively), but it may be premature to draw firm conclusions on the basis of the HN sequence alone. When data are available from other genes such relationships may become clearer.

We thank Kate Brown and Tony Pickard for technical assistance. This work was supported by grants from the Wellcome Trust and North of England Cancer Research Campaign. Neil Millar is supported by a Science and Engineering Research Council studentship.

REFERENCES


NDV HN gene sequence


*Received 1 May 1986*