

# A $^1\text{H}$ NMR database computer program for the analysis of the primary structure of complex carbohydrates

J. Albert van Kuik, Karl Hård and Johannes F.G. Vliegenthart

*Bijvoet Center, Department of Bio-Organic Chemistry, Utrecht University, P.O. Box 80.075,  
3508 TB Utrecht (Netherlands)*

(Received January 9th, 1992; accepted April 7th, 1992)

## ABSTRACT

A  $^1\text{H}$  NMR database computer program has been developed to determine the primary structure of complex carbohydrates. The database contains carbohydrate structures, their corresponding  $^1\text{H}$  NMR data, and literature references. From an input list of chemical shift values, the program generates an output list of partially or completely matching carbohydrate structures. In order to facilitate the recognition of the matching part of the selected carbohydrate structures, these structures are displayed with the matching structural elements highlighted. This new  $^1\text{H}$  NMR database, together with the search program described, now provides a fast access to the published  $^1\text{H}$  NMR data of complex carbohydrates and furnishes easy links to carbohydrate structures. The performance of the program is demonstrated by the analysis of five carbohydrate fractions prepared from a pool of horse serum glycoproteins.

## INTRODUCTION

High-resolution  $^1\text{H}$  NMR spectroscopy is now in common use for the determination of the structure of complex carbohydrates and the amount of data is ever growing. Although compilations of these data appear from time to time<sup>1,2</sup>, continual updating and evaluation are necessary. Therefore, it is desirable to store, update, and evaluate such NMR data on a computerised basis and this approach is being developed in various laboratories. Direct analyses (a) of spectra<sup>3,4</sup>, (b) of tables of assigned or unassigned chemical shift data<sup>5,6</sup>, and (c) by comparison of measured and simulated spectra<sup>7,8</sup> have been explored, but no practical NMR database of complex carbohydrate structures has emerged. An ideal solution would be a database of all published  $^1\text{H}$  and  $^{13}\text{C}$  NMR spectra of complex carbohydrates, but such a compilation is difficult, even if the various laboratories donate their original data. Moreover, many published  $^1\text{H}$  NMR spectra of biologically interest-

---

*Correspondence to:* Professor J.F.G. Vliegenthart, Bijvoet Center, Department of Bio-Organic Chemistry, Utrecht University, P.O. Box 80.075, 3508 TB Utrecht, Netherlands.

ing carbohydrates are of poor quality, due to the small amounts of material available, the presence of mixtures of different carbohydrates, or the occurrence of non-carbohydrate impurities. Usually, small amounts of carbohydrate material also imply that no  $^{13}\text{C}$  NMR spectra are available. Moreover,  $^1\text{H}$  NMR spectra recorded at different frequencies have different appearances.

We now report on the development of a database of  $^1\text{H}$  NMR chemical shifts, together with a computer program for easy access. The program is demonstrated by the analysis of five fractions of *N*-linked oligosaccharides isolated from horse serum glycoproteins.

## EXPERIMENTAL

*The  $^1\text{H}$  NMR database program.*—The program is written in language C and runs on IBM compatible Personal Computers, using MS DOS (version 3.3 or higher). Currently, the program and database require 3.5 Mbytes of disc space. The database combines  $^1\text{H}$  NMR data taken from the literature with complex carbohydrate structures and data taken from the Complex Carbohydrate Structure Database<sup>9</sup> (CCSD). All chemical shifts and monosaccharide residues are indexed in order to improve the search performance. At present, the database contains 508 tables of  $^1\text{H}$  NMR chemical shifts selected from 102 publications with an emphasis on carbohydrate chains released from glycoproteins. Only  $^1\text{H}$  NMR data for solutions in  $\text{D}_2\text{O}$  are included in the database and all chemical shifts are referenced to the signal of acetone<sup>1</sup> (2.225 ppm). The tables can be presented in ASCII text format and a typical output is depicted in Table I.

In order to search the database, an input list of chemical shifts is required. Either correct chemical shifts can be derived from doublets and multiplets, or an average value for each multiplet may be entered. Although coupling constants can be stored in the database, this information is not used by the current version of the program.

Before a search can be initiated, the following three parameters have to be set (see Table IIA).

(1) *The tolerance limit of the values of the chemical shifts.* Usually a value of 0.005 ppm is appropriate.

(2) *The threshold for a hit.* This value is expressed as the percentage of matching chemical shift values between the search list and the database. The selection of this value depends on the nature of the sample (single compound or a mixture) and by the intention to search either for complete or partial structures.

(3) *The type of carbohydrate chain.* The structures in the database are grouped, according to the type of carbohydrate chain, into four sections which can be accessed separately, namely, lactose-type [i.e., carbohydrate chains containing a  $\beta\text{-D-Galp-(1}\rightarrow\text{4)-}\beta\text{-D-Glcp}$  sequence], *N*-linked, *O*-linked, and polysaccharides. The main effort so far has been focused on *N*- and *O*-linked carbohydrate chains of glycoproteins. Even if the type of linkage is known, it may be useful to select

TABLE I

An example of a  $^1\text{H}$  NMR database record presented in ASCII format

H#: N-0B02-003873

CC: CCSD:3873

AU: Damm JBL; Voshol H; Hard K; Kamerling JP; Vliegenthart JFG

JL: Eur. J. Biochem. (1989) 180: 101–110

SC: N2.2A

$\alpha\text{-D-Neup5Ac-(2-6)-}\beta\text{-D-Galp-(1-4)-}\beta\text{-D-GlcpNAc-(1-2)-}\alpha\text{-D-Manp-(1-6)}$   
 $\beta\text{-D-Manp-(1-4)-}\beta\text{-D-GlcpNAc-(1-4)-D-GlcNAc}$   
 $\alpha\text{-D-Neup4Ac5Ac-(2-6)-}\beta\text{-D-Galp-(1-4)-}\beta\text{-D-GlcpNAc-(1-2)-}\alpha\text{-D-Manp-(1-3)}$

MHz 500  
 TEMP 300  
 SOLV D20

Residue	Linkage	Proton	PPM	J	Hz	Note
D-GlcNAc		H-1 $\alpha$	5.190			
		H-1 $\beta$	4.694			
		NAc	2.038			
$\beta\text{-D-GlcpNAc}$	4	H-1( $\alpha$ )	4.614			
		H-1( $\beta$ )	4.606			
		NAc	2.085			
$\beta\text{-D-Manp}$	4,4	H-1	4.77			a
		H-2	4.260			
$\alpha\text{-D-Manp}$	6,4,4	H-1	4.950			
		H-2	4.117			
$\beta\text{-D-GlcpNAc}$	2,6,4,4	H-1	4.606			
		NAc	2.066			
$\beta\text{-D-Galp}$	4,2,6,4,4	H-1	4.447			
$\alpha\text{-D-Neup5Ac}$	6,4,2,6,4,4	H-3ax	1.719			
		H-3eq	2.673			
		H-4	3.660			a
		NAc	2.030			
$\alpha\text{-D-Manp}$	3,4,4	H-1	5.143			
		H-2	4.204			
$\beta\text{-D-GlcpNAc}$	2,3,4,4	H-1	4.606			
		NAc	2.110			
$\beta\text{-D-Galp}$	4,2,3,4,4	H-1	4.440			
$\alpha\text{-D-Neup4Ac5Ac}$	6,4,2,3,4,4	H-3ax	1.852			
		H-3eq	2.677			
		H-4	4.904			
		H-5	4.052			a
		NAc	1.964			
		OAc	2.077			

NOTE a) Value obtained at 310 K.

more than one type for a search, because some structural elements occur in more than one type of chain.

After setting the parameters 1–3, a list of chemical shifts can be entered. If the monosaccharide composition of the sample is known, an optional list of residue

TABLE II

Examples of input and output screens of the program

## A Input screen

File Edit Search Range Options Window Help

Search for PPM

> 0.005 (Tolerance limit in ppm)

> 50% (Matching ppm's needed for a hit)

> N [L,N,O,P] Lactose-type, N-,O-linked, Poly.

PPM VALUES

5.190

5.143

4.958

4.925

4.901

4.613

4.609

4.604

4.574

4.552

4.444

4.440

4.256

File Help Line: 0 Column: 0 5.19ppm

B Output screen which presents structures and the corresponding  $^1\text{H}$  NMR data <sup>a</sup>

File Edit Search Range Options Window Help

SW: N-0A02-003860

B-D-Galp-(1-4)-B-D-GlcpNAc-(1-2)-B-D-Manp-(1-6)

B-D-Manp-(1-3)-B-D-Galp-(1-4)-B-D-GlcpNAc-(1-2)-B-D-Manp-(1-3)

HW: N-0A02-003860 CC: CCSD:3860 MHz 500 Temp 300 Solv D2O

Residue	Linkage	Proton PPM	J	Hz	No
D-GlcNAc		H-1 $\alpha$ 5.189			
		NAc 2.038			
B-D-GlcpNAc	4	H-1( $\alpha$ ) 4.614			
		H-1( $\beta$ ) 4.606			
		NAc 2.079			
B-D-Manp	4,4	H-2 4.255			
$\alpha$ -D-Manp	6,4,4	H-1 4.930			
		H-2 4.110			

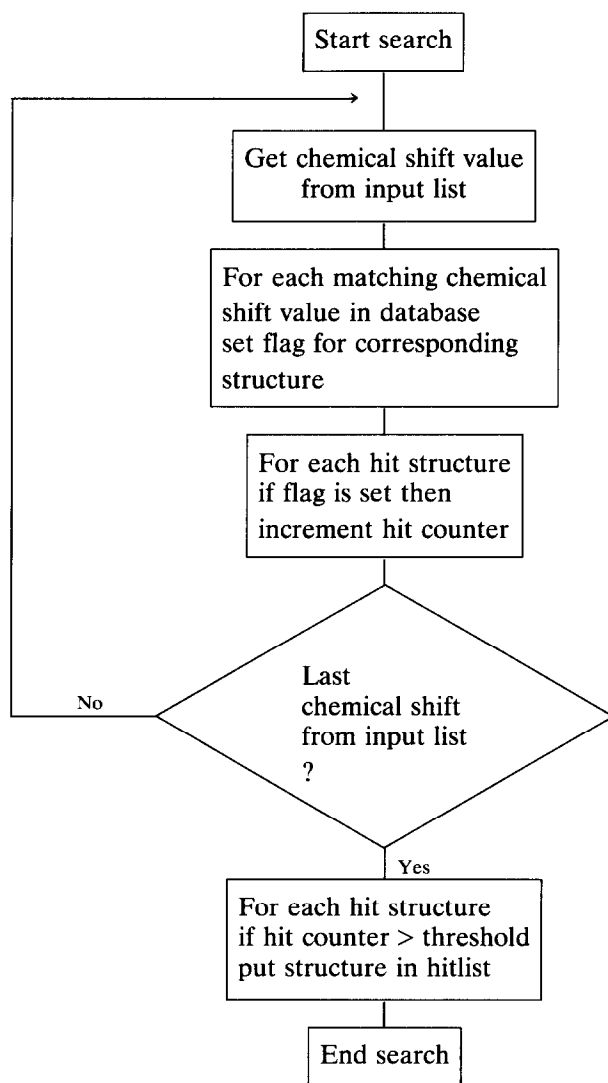
File Help 118 ppm hits = 84 % match 5.19ppm

<sup>a</sup> The matching residues and  $\delta$  values are highlighted.

constraints may be added to the search profile. Only chemical shift data for the relevant residues will then be used in the search. The program performs a straightforward indexed search whereby all chemical shifts inside the tolerance limit and belonging to residues from the correct type of carbohydrate chain (i.e., hits) are counted. If, at the end of the search, a structure has enough hits to meet the threshold, it is placed in a hitlist (see Scheme 1 for the search strategy). Hitlists from different searches can also be combined using 'AND', 'OR', or 'NOT' operations. When a hitlist is examined, the structures and the corresponding  $^1\text{H}$  NMR data are displayed. In the structures, the residues with matching chemical shift data are highlighted, as are the matching chemical shifts in the NMR data (Table IIB). The output can also be directed to a printer whereby  $^1\text{H}$  NMR data, structures, and literature references are combined (Table I). Another way of searching the database is to look for structures that match with a user-defined profile of structures.

*Preparation of oligosaccharides.*—In order to test the database program, a collection of *N*-linked oligosaccharides prepared from horse serum glycoproteins was used. The glycoproteins were isolated by precipitation with sulfosalicylic acid from 5 L of horse blood according to Sander et al.<sup>10</sup>. The *N*-linked carbohydrate chains were released enzymically using peptide-*N*<sup>4</sup>-(*N*-acetyl- $\beta$ -D-glucosaminy)-asparagine amidase F (PNGase F) according to a modified version of a described protocol<sup>11</sup>. The sample (400 mg) of glycoprotein was dissolved in 50 mM Tris (40 mL, adjusted to pH 7.2 with HCl, and containing 50 mM EDTA, 1% of sodium dodecyl sulfate, 0.2% of 2-mercaptoethanol, and 0.02% of sodium azide). The solution was incubated at 40° for 1 h, the non-ionic detergent Nonidet P-40 was added to 2%, and the solution was incubated at room temperature with several batches of PNGase-F (a total of 73 U during 6 days). The carbohydrate chains released were separated from the *N*-deglycosylated protein by gel-permeation chromatography on a column (1.8  $\times$  48 cm) of Bio-Gel P-100 by elution with 50 mM  $\text{NH}_4\text{HCO}_3$ , adjusted to pH 7.2 with HCl. Carbohydrate-containing fractions were combined, lyophilised, and desalted on a column (1.2  $\times$  20 cm) of Bio-Gel P-2 by elution with water. The carbohydrate chains were fractionated according to charge on a Mono Q HR 5/5 anion-exchange column (Pharmacia FPLC system), using a NaCl gradient essentially as described<sup>11</sup>. This fractionation gave rise to three carbohydrate-positive peaks, denoted N1–N3 (chromatogram not shown), which had the elution volumes of monosialylated diantennary, disialylated diantennary, and trisialylated triantennary reference compounds, respectively. The carbohydrate-positive Mono Q fractions were sub-fractionated by HPLC on a Lichrosorb-NH<sub>2</sub> column (0.46  $\times$  25 cm, Chrompack) by elution with 35:65 30 mM potassium phosphate (pH 7.0)–acetonitrile at 2 mL/min. The fractions were desalted on Bio-Gel P-2.

*$^1\text{H}$  NMR spectroscopy.*—Each HPLC fraction was treated repeatedly with D<sub>2</sub>O at room temperature and lyophilised after each treatment. Finally, each sample was redissolved in D<sub>2</sub>O (0.45 mL, 99.96 atom% D, Aldrich).  $^1\text{H}$  NMR spec-



Scheme 1. Flow diagram of the search strategy of the  $^1\text{H}$  NMR database program.

troscopy (600 MHz) was performed with a Bruker AM-600 spectrometer (Department of Biophysical Chemistry, Nijmegen University). Chemical shifts ( $\delta$ ) are referenced to internal sodium 4,4-dimethyl-4-silapentane-1-sulfonate, but were measured by reference to internal acetone ( $\delta$  2.225 at  $27^\circ$ ).

## RESULTS

In order to test the  $^1\text{H}$  NMR database computer program, five carbohydrate-containing HPLC fractions prepared from horse serum glycoproteins were anal-

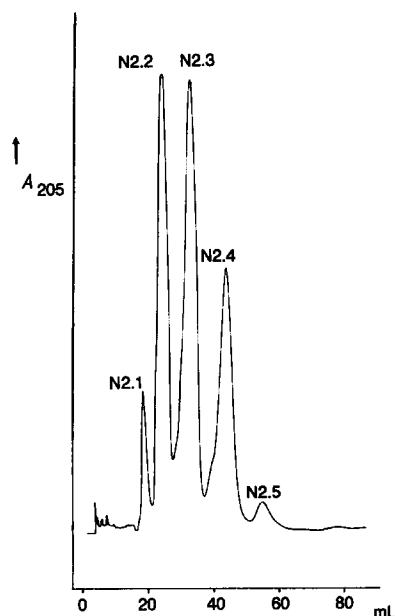


Fig. 1. Sub-fractionation of FPLC fraction N2 by HPLC on a Lichrosorb-NH<sub>2</sub> column (0.46 × 25 cm, Chrompack).

ysed. These fractions were derived from FPLC fraction N2 (see Experimental). Based on its behaviour on Mono Q, fraction N2 contained compounds with two negative charges. HPLC of fraction N2 yielded five fractions, denoted N2.1–N2.5 (Fig. 1). Because of the similarity of the compounds in these fractions, they provide a suitable test for the computer program.

Fractions N2.1–N2.5 contain only *N*-linked carbohydrate chains, since they were released from the glycoproteins by PNGase-F. Therefore, only the chemical shift data in the 265 tables for *N*-linked carbohydrate chains were searched (tolerance limit set to 0.005 ppm). For each component, only the chemical shifts of the signals in the regions for the so-called structural reporter group were searched, namely, the signals in the ranges 5.6–4.0 and 3.5–1.0 ppm. The signals close to those of water (4.7–4.8 ppm), acetone (2.225 ppm), and acetate (1.908 ppm), and a commonly occurring non-carbohydrate doublet at 1.32 ppm, were excluded.

The identification of the fractions is described in order of decreasing retention time in HPLC. The chemical shift data for the compounds in fractions N2.1–N2.5 are listed in Table III. The numbering of the monosaccharide residues of the carbohydrate chains is shown in Fig. 2. The results of the search for each fraction are depicted in Table IV.

**Fraction N2.5.**—The region for NAc signals (not shown) indicated that fraction N2.5 contains a mixture of carbohydrate chains. The signals of the minor components of this fraction (10%) are not discussed. From the equal intensity of the H-1

TABLE III

<sup>1</sup>H NMR chemical shifts <sup>a</sup> of constituent monosaccharides for the diantennary carbohydrate chains derived from HPLC fractions N2.1–N2.5








Residue <sup>b</sup>	Proton(s)	N2.1 <sup>c</sup>	N2.2	N2.3a	N2.3b	N2.4	N2.5a	N2.5b
								
GlcNAc-1	H-1 $\alpha$	5.190	5.189	5.189	5.189	5.190	5.190	5.190
	H-1 $\beta$	4.696	4.695	4.695	4.695	4.695	4.695	4.695
	NAc	2.039	2.038	2.038	2.038	2.038	2.038	2.038
GlcNAc-2 <sup>e</sup>	H-1( $\alpha$ )	4.613	4.615	4.616	4.616	4.616	4.616	4.616
	H-1( $\beta$ )	4.604	4.609	4.606	4.606	4.605	4.605	4.605
	NAc( $\alpha/\beta$ )	2.082	2.085	2.084	2.084	2.084	2.084	2.084
	NAc( $\beta$ )		2.083					
Man-3	H-1	n.d. <sup>f</sup>	4.780	4.778	4.778	4.776	n.d.	n.d.
	H-2	4.256	4.259	4.258	4.258	4.255	4.255	4.255
Man-4'	H-1	4.925	4.956	4.948	4.955	4.947	4.947	4.947
	H-2	4.119	4.124	4.120	4.120	4.117	4.119	4.119
GlcNAc-5'	H-1	4.574	4.609	4.606	4.606	4.605	4.605	4.605
	NAc	2.043	2.103	2.065	2.103	2.065	2.069	2.066
Gal-6'	H-1	4.552	4.444	4.447	4.443	4.448	4.452	4.448
	H-3	4.134	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Neu4,5Ac <sub>2</sub> '	H-3 <sub>ax</sub>	1.927	1.849		1.849			
	H-3 <sub>eq</sub>	2.772	2.681		2.681			
	H-4	4.958	4.904		4.903			
	H-5	4.090	4.053		4.053			
	OAc	2.073	2.077		2.077			
	NAc	1.964	1.964		1.964			
Neu5Ac'	H-3 <sub>ax</sub>			1.717		1.717		1.717
	H-3 <sub>eq</sub>			2.676		2.674		2.677
	NAc			2.030		2.030		2.030
Neu5Gc'	H-3 <sub>ax</sub>						1.734	
	H-3 <sub>eq</sub>						2.691	
	NGc						4.119	
Man-4	H-1	5.143	5.144	5.143	5.135	5.134	5.135	5.135
	H-2	4.202	4.202	4.200	4.200	4.197	4.197	4.197
GlcNAc-5	H-1	4.609	4.609	4.606	4.606	4.605	4.605	4.605
	HAc	2.109	2.109	2.109	2.069	2.069	2.069	2.072
Gal-6	H-1	4.440	4.440	4.440	4.443	4.444	4.443	4.448
Neu4,5Ac <sub>2</sub>	H-3 <sub>ax</sub>	1.849	1.849	1.849				
	H-3 <sub>eq</sub>	2.677	2.677	2.676				
	H-4	4.901	4.901	4.903				
	H-5	4.053	4.053	4.053				
	OAc	2.076	2.077	2.077				
	NAc	1.964	1.964	1.964				



TABLE III (continued)

Residue <sup>b</sup>	Proton(s)	N2.1 <sup>c</sup> 4N <sup>d</sup> — 4N <sup>d</sup> —	N2.2 4N <sup>d</sup> — 4N <sup>d</sup> —	N2.3a N <sup>d</sup> — 4N <sup>d</sup> —	N2.3b 4N <sup>d</sup> — N <sup>d</sup> —	N2.4 N <sup>d</sup> — N <sup>d</sup> —	N2.5a G <sup>d</sup> — N <sup>d</sup> —	N2.5b N <sup>d</sup> — G <sup>d</sup> —
Neu5Ac	H-3 <sub>ax</sub>				1.717	1.717	1.717	
	H-3 <sub>eq</sub>				2.667	2.668	2.669	
	NAc				2.030	2.030	2.030	
Neu5Gc	H-3 <sub>ax</sub>							1.734
	H-3 <sub>eq</sub>							2.683
	NGc							4.119

<sup>a</sup> Chemical shifts are relative to acetone at 2.225 ppm in D<sub>2</sub>O at 300 K, acquired at 600 MHz. <sup>b</sup> For the numbering of the residues, see Fig. 2. <sup>c</sup> N =  $\alpha$ -D-Neup5Ac-(2  $\rightarrow$  6), 4N =  $\alpha$ -D-Neup4,5Ac<sub>2</sub>-(2  $\rightarrow$  3), 4N =  $\alpha$ -D-Neup4,5Ac<sub>2</sub>-(2  $\rightarrow$  6), G =  $\alpha$ -D-Neup5Gc-(2  $\rightarrow$  6). <sup>d</sup> The diantennary structure  $\beta$ -D-Galp-(1  $\rightarrow$  4)- $\beta$ -D-GlcpNAc-(1  $\rightarrow$  2)- $\alpha$ -D-Manp-(1  $\rightarrow$  6)-[ $\beta$ -D-Galp-(1  $\rightarrow$  4)- $\beta$ -D-GlcpNAc-(1  $\rightarrow$  2)- $\alpha$ -D-Manp-(1  $\rightarrow$  3)]- $\beta$ -D-Manp-(1  $\rightarrow$  4)- $\beta$ -D-GlcpNAc-(1  $\rightarrow$  4)-D-GlcNAc is represented by the symbolic notation  $\left[ \begin{array}{c} \text{N} \\ \text{4N} \end{array} \right] -$ . <sup>e</sup>  $\alpha$  and  $\beta$  in parentheses refer to the anomeric configuration of GlcNAc-1. <sup>f</sup> Not determined.

signals of the major components, a mixture of carbohydrate chains in equal proportions is expected. In order to search the database, 19 chemical shifts were selected from the spectrum and the threshold was set to 80% matching, which resulted in the 4 hits depicted in Table IV. Each of the hits was a disialylated diantennary compound with a sialic acid residue that was either Neup5Ac or Neup5Gc. From the <sup>1</sup>H NMR spectrum alone, it was not clear whether fraction N2.5 contained all four theoretically possible carbohydrate chains, or only two compounds with one Neup5Ac and one Neup5Gc residue each. However, since carbohydrate chains with different sialic acid residues are separated by HPLC under the conditions applied, the presence of carbohydrate chains with either two Neup5Ac or two Neup5Gc residues is excluded. Therefore, fraction N2.5 contains the two compounds each with one Neup5Ac and one Neup5Gc residue.

**Fraction N2.4.**—The <sup>1</sup>H NMR spectrum of fraction N2.4 (not shown) revealed a single component, and 18 chemical shifts were selected and used in the search (> 80% match). An output of 9 closely related disialylated diantennary structures was obtained (Table IV). One structure contained two  $\alpha$ -D-Neup5Ac-(2  $\rightarrow$  6) residues and gave a 100% match. A manual comparison of the <sup>1</sup>H chemical shift data for fraction N2.4 with all the values from this 100%-match table confirmed that the structure of fraction N2.4 and the hit structure were identical. Inspection of the data of the other eight hit structures showed the mismatch to be in the chemical shifts of the  $\alpha$ -D-Neup5Ac-(2  $\rightarrow$  3),  $\alpha$ -D-Neup4,5Ac<sub>2</sub>-(2  $\rightarrow$  6), and/or  $\alpha$ -D-Neup5Gc-(2  $\rightarrow$  6) residues, or of GlcNAc-1 and GlcNAc-2, due to the presence of a Fuc residue in some structures with a lower match. This manual check of the selected structure(s) is vital because the program can make erroneous assignments when coupling constants are not utilized. In this example, H-2 of Man-4' resonates

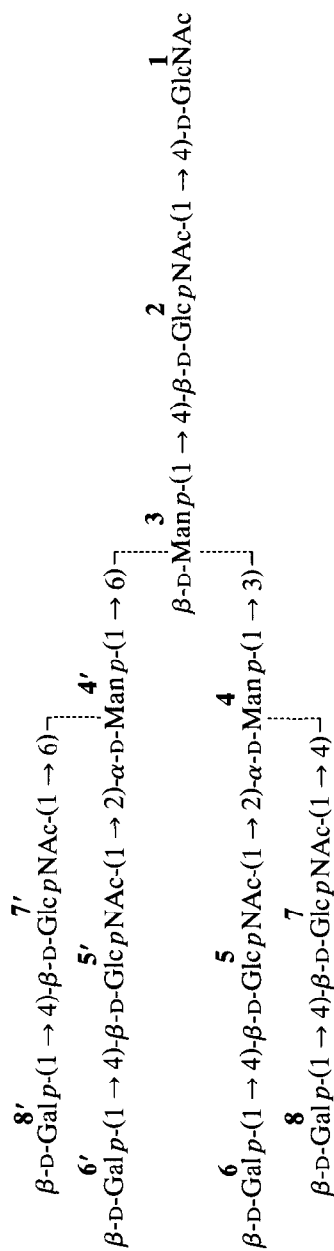
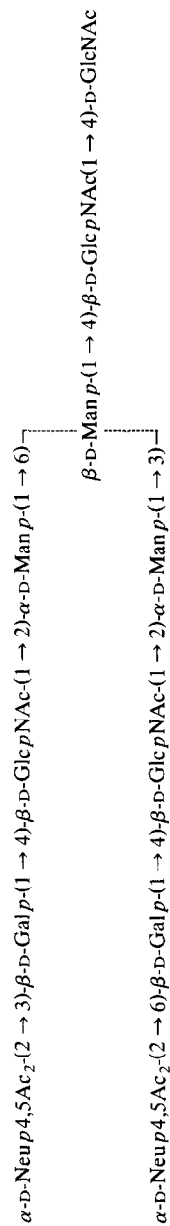


Fig. 2. Numbering of the constituent monosaccharides in the carbohydrate chains.



Structure 1

TABLE IV

Performance of the program on the identification of five HPLC-purified carbohydrate fractions, derived from a pool of horse serum glycoproteins

Fraction code	Search		Result			
	Number of input chemical shifts	Threshold (%)	Search <sup>a</sup> time (s)	Number of hits	Match (%)	Hit structure <sup>b,c</sup>
N2.5	19	80	8	4	89	N-6'-5'-4' } 3-2-1
						N-6-5-4 } N-6'-5'-4' } 3-2-1
					100	G-6-5-4 } G-6'-5'-4' } 3-2-1
						N-6-5-4 } G-6'-5'-4' } 3-2-1
					84	G-6'-5'-4' } 3-2-1 G-6-5-4 }
N2.4	18	80	9	9	83	5'-4' } 3-2-1 N-6-5-4 }
					88	6'-5'-4' } 3-2-1 N-6-5-4 }
						N-6'-5'-4' } 3-2-1 4N-6-5-4 }
					83	4N-6'-5'-4' } 3-2-1 N-6-5-4 }
						N-6'-5'-4' } 3-2-1 N-6-5-4 }
					100	N-6'-5'-4' } 3-2-1 N-6-5-4 }
						N-6'-5'-4' } 3-2-1 N-6-5-4 }
					88	N-6-5-4 } 3-2-1 N-6'-5'-4' }
						N-6-5-4 } 3-2-1 N-6'-5'-4' }
					94	G-6-5-4 } 3-2-1 G-6'-5'-4' }
						G-6-5-4 } 3-2-1 G-6'-5'-4' }
					83	N-6'-5'-4' } 3-2-1 N-6-5-4 }
						N-6'-5'-4' } 3-2-1 N-6-5-4 }
N2.3	23	80	9	2	95	N-6'-5'-4' } 3-2-1 4N-6-5-4 }
						4N-6'-5'-4' } 3-2-1 N-6-5-4 }
					95	N-6'-5'-4' } 3-2-1 N-6-5-4 }

TABLE IV (continued)

Fraction code	Search		Result			
	Number of input chemical shifts	Threshold (%)	Search <sup>a</sup> time (s)	Number of hits	Match (%)	Hit structure <sup>b,c</sup>
N2.2	21	80	6	4	100	<b>4N-6'-5'-4'</b> } 3-2-1
						<b>4N-6-5-4</b> } N-6'-5'-4' } 3-2-1
						<b>4N-6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1
						<b>N-6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1-Asn
						<b>4N-6-5-4</b> }
N2.1	28	50	11	12	64	<b>6'-5'-4'</b> } 3-2-1
						<b>4N-6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1
						<b>4N-6-5-4</b> } <b>N-6'-5'-4'</b> } 3-2-1
						<b>4N-6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1
						<b>N-6-5-4</b> } <b>N-6'-5'-4'</b> } 3-2-1
						<b>N-6-5-4</b> } <b>N-6'-5'-4'</b> } 3-2-1
						<b>N-6-5-4</b> } <b>N-6'-5'-4'</b> } 3-2-1
						<b>G-6-5-4</b> } <b>N-6'-5'-4'</b> } 3-2-1
						<b>G-6-5-4</b> } <b>6'-5'-4'</b> } 3-2-1-Asn
						<b>4N-6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1-Asn
						<b>6-5-4</b> } <b>4N-6'-5'-4'</b> } 3-2-1-Asn
						<b>4N-6-5-4</b> } <b>N-8'-7'</b> } <b>N-6'-5'-4'</b> } 3-2
						<b>N-6-5-4</b> } <b>N-8-7</b> }

<sup>a</sup> Measured on a 20-MHz IBM-PS/2 Model 80 computer, using a 1-Mb disk cache. <sup>b</sup> For the numbering of the monosaccharide residues, see Fig. 2. N =  $\alpha$ -D-Neup5Ac-(2  $\rightarrow$  6),  $\bar{N}$  =  $\alpha$ -D-Neup5Ac-(2  $\rightarrow$  3), 4N =  $\alpha$ -D-Neup4,5Ac<sub>2</sub>-(2  $\rightarrow$  6), G =  $\alpha$ -D-Neup5Gc-(2  $\rightarrow$  6), F =  $\alpha$ -L-Fucp-(1  $\rightarrow$  6). <sup>c</sup> Hit structures that entirely match the components in the fractions N2.1–N2.5 are printed in bold.

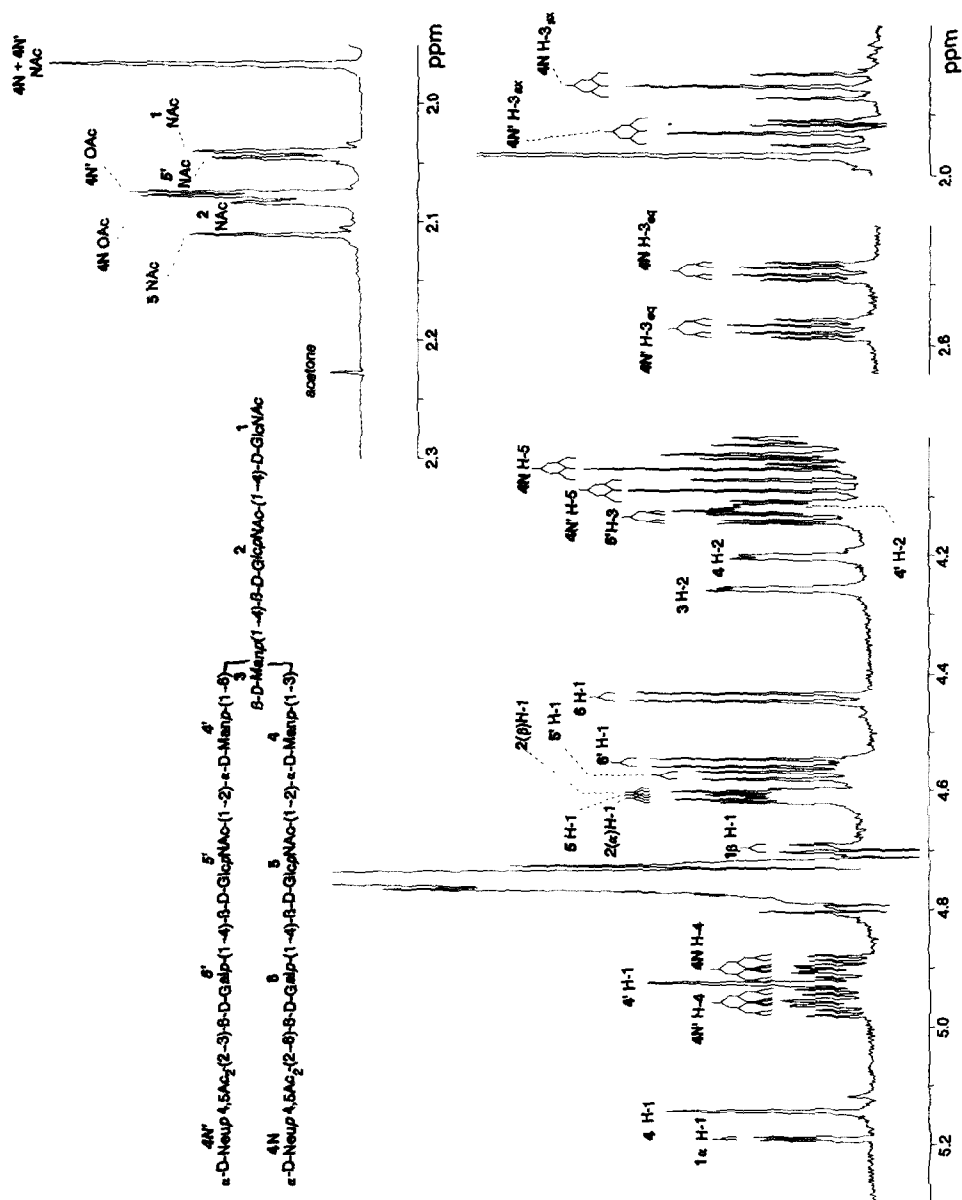


Fig. 3.  $^1\text{H}$  NMR spectrum (600 MHz) of the structural reporter group regions of the oligosaccharide N2.1.

at the same position as the glycolyl methylene protons of Neu $p$ 5Gc in the compounds containing Neu $p$ 5Gc, and the program gives a false match for the latter signal.

*Fraction N2.3.*—The  $^1\text{H}$  NMR spectrum of fraction N2.3 (not shown) indicated the presence of a mixture of closely related components. A search with an input list of 23 chemical shifts and an  $> 80\%$  match resulted in an output of two disialylated diantennary structures, each of which contained one Neu $p$ 5Ac and one Neu $p$ 4,5Ac $_2$  residue (Table IV). Each hit structure gave a 95% match. A manual comparison of the hit NMR data with the  $^1\text{H}$  NMR spectrum confirmed that fraction N2.3 contained the two carbohydrate chains that corresponded with these hit structures.

*Fraction N2.2.*—The  $^1\text{H}$  NMR spectrum of fraction N2.2 (not shown) indicated the presence of a pure compound. A search with 21 chemical shifts ( $> 80\%$  match) resulted in 4 hits (Table IV). One hit structure had a 100% match and was a disialylated diantennary oligosaccharide with two Neu $p$ 4,5Ac $_2$  residues. Two other hit structures, each with a lower percentage match, corresponded to the structures found in the search for fraction N2.3, but were not present in fraction N2.2. The output data revealed that the Neu $p$ 5Ac residues, present in these two lower-match structures, were not present in fraction N2.2. The presence of the third structure with a lower match percentage was excluded because it contained an Asn residue, which has a profound effect on the chemical shift data for GlcNAc-1. All chemical shifts for the 100% match could be located in the spectrum of fraction N2.2 and proved that the hit structure was correct.

*Fraction N2.1.*—The  $^1\text{H}$  NMR spectrum of fraction N2.1 (see Fig. 3) revealed a single component. A search using 28 chemical shifts ( $> 80\%$  match) resulted in no hits. Hence, the  $^1\text{H}$  NMR data for this carbohydrate chain are not in the database. When the threshold percentage was lowered to 50%, a search produced 12 hits (Table IV), most of which were diantennary compounds resembling those discussed above, and one tetra-antennary compound. Structures with Neu $p$ 4,5Ac $_2$  residues had the highest percentage match. Comparison of the chemical shifts of the input list and the hit  $^1\text{H}$  NMR data indicated that two chemical shifts (1.927 and 2.772 ppm) in the former had no matching value in the latter. A search with these two chemical shifts of all the types of carbohydrate chain, with a 100% match, resulted in one hit structure, namely,  $\alpha$ -D-Neu $p$ 4,5Ac $_2$ -(2  $\rightarrow$  3)- $\beta$ -D-Galp-(1  $\rightarrow$  4)-D-Glc $^{12}$ . This structure contains a Neu $p$ 4,5Ac $_2$  residue in a (2  $\rightarrow$  3)- $\alpha$  linkage, instead of the (2  $\rightarrow$  6)- $\alpha$  linkage. Subsequent inspection of the spectrum of N2.1 indicated that two  $\alpha$ -Neu $p$ 4,5Ac $_2$  residues formed part of the carbohydrate chain of fraction N2.1, with one (2  $\rightarrow$  6)-linked and the other one (2  $\rightarrow$  3)-linked to  $\beta$ -D-Galp. The H-1 signal of Man-4 at 5.143 ppm revealed  $\alpha$ -D-Neu $p$ 4,5Ac $_2$ -(2  $\rightarrow$  6) to be located in the (1  $\rightarrow$  3)-linked branch, and the H-1 signal of Man-4' at 4.925 ppm confirmed the presence of  $\alpha$ -D-Neu $p$ 4,5Ac $_2$ -(2  $\rightarrow$  3) in the (1  $\rightarrow$  6)-linked branch $^1$ , and indicated a novel compound with the structure 1.

## DISCUSSION

The increasing number of published structures of oligosaccharides and the corresponding amount of NMR data make it difficult to keep track of all of the parameters needed for the analysis of a structure. Therefore, the accumulation of these data in computerised databases is indispensable. The construction of a  $^1\text{H}$  NMR database has a high priority since the use of such data enables the unambiguous and non-destructive identification of carbohydrate structures. In the present report, the functioning of a  $^1\text{H}$  NMR database of complex carbohydrate structures and its corresponding management program are described. The program, which searches the database with an input list of chemical shifts, can discriminate between closely related carbohydrates. If the entry list contains chemical shift data for a mixture of carbohydrates, the program gives an output list of possible structures which have to be checked manually. The highlighted presentation of the matching chemical shifts and residues is helpful in this check. Even if the compound under investigation does not have a matching structure in the database, the program can aid in the determination of the structure by giving examples of structural elements that might be present.

In the analysis of fractions N2.1–N2.5, only one of the many possible search strategies has been discussed. The use of a smaller number of chemical shifts in a search normally gives a broader range of matching structures. The optimal number of chemical shifts, which is not known in advance, depends on the structure, the quality of the spectrum, and the purity of the sample. Usually, a search is started with a small number of chemical shifts which is increased gradually, thereby narrowing the range of possible matching structures. No use has been made of the option to employ the names of residues in a search, nor has the option to search for structures with a structure–search profile been illustrated. These options could be useful for combining  $^1\text{H}$  NMR data with data obtained, for example, from monosaccharide and methylation analysis. An option to expand the search information with coupling constants, or knowledge of multiplet patterns, will make the search more discriminating in the acceptance of chemical shifts.

The features of the program enable known carbohydrate structures to be identified on a routine basis. Additionally, the program can assist in the analysis of  $^1\text{H}$  NMR spectra of unknown carbohydrate chains. Often, additional information from other sources will be needed for unambiguous identification of the structures. By storing complete structures exclusively in the database, the mutual influence of connected structural elements on chemical shifts is implicit. The program and database are designed to grow and to be linked easily to other computer platforms. The presence of a careful selection of relevant reference  $^1\text{H}$  NMR data in the database is the key for a good performance of the program and determines its value as a tool for the analysis of carbohydrate structures.

## ACKNOWLEDGMENTS

We thank Professor Dr. G.J.W. van der Meij (Department of Veterinary Medicine, Utrecht University) for his kind donation of the horse serum, and J.J.G. van Soest for experimental help. This work was supported by grants from the EC Biotechnology Action Program BAP-0364-NL, the EC Biotechnology Research for Innovation, Development and Growth in Europe (BRIDGE) BIOT-CT90-0184, and the Netherlands Foundation for Chemical Research (SON/NWO). Support (to K.H.) in part by the Finnish Cultural Foundation and the Magnus Ehrnrooth Foundation is acknowledged.

## REFERENCES

- 1 J.F.G. Vliegthart, L. Dorland, and H. van Halbeek, *Adv. Carbohydr. Chem. Biochem.*, 41 (1983) 209–374.
- 2 J.P. Kamerling and J.F.G. Vliegthart, *Biol. Magn. Reson.*, 10 (1992) 1–194.
- 3 J. Thomsen and B. Meyer, *J. Magn. Reson.*, 84 (1989) 212–217.
- 4 B. Meyer, T. Hansen, D. Nute, P. Albersheim, A. Darvill, W. York, and J. Sellers, *Science*, 251 (1991) 542–544.
- 5 D.S.M. Bot, P. Cleij, H.A. van 't Klooster, H. van Halbeek, G.A. Veldink, and J.F.G. Vliegthart, *J. Chemometrics*, 2 (1988) 11–27.
- 6 E.F. Hounsell and D.J. Wright, *Carbohydr. Res.*, 205 (1990) 19–29.
- 7 P.-E. Jansson, L. Kenne, and G. Widmalm, *Carbohydr. Res.*, 193 (1989) 322–325.
- 8 H. Baumann, P.-E. Jansson, L. Kenne, and G. Widmalm, *Carbohydr. Res.*, 221 (1991) 183–190.
- 9 S. Doubet, K. Bock, D. Smith, A. Darvill, and P. Albersheim, *Trends Biochem. Sci.*, 14 (1989) 475–477.
- 10 M. Sander, R.W. Veh, and R. Schauer, *Proc. Int. Symp. Glycoconjugates, 5th*, Kiel, 1979, pp 358–359.
- 11 K. Hård, A. Mekking, J.B.L. Damm, J.P. Kamerling, W. de Boer, R.A. Wijnands, and J.F.G. Vliegthart, *Eur. J. Biochem.*, 193 (1990) 263–271.
- 12 J.P. Kamerling, L. Dorland, H. van Halbeek, J.F.G. Vliegthart, M. Messer, and R. Schauer, *Carbohydr. Res.*, 100 (1982) 331–340.