Carbohydrate Research, 240 (1993) 57–69
Elsevier Science Publishers B.V., Amsterdam

The application of NMR–pattern-recognition methods to the classification of reduced, peracetylated oligosaccharide residues

Warren J. Goux and Denise S. Weber

Department of Chemistry, University of Texas at Dallas, Richardson, TX 75083-0688 (USA)

(Received February 25th, 1992; accepted September 22nd, 1992)

ABSTRACT

In the present paper homo- and hetero-nuclear correlation spectroscopies have been used to assign proton and carbonyl carbon resonances of a number of reduced, peracetylated mono- and oligo-saccharide derivatives. Each of the native structures for which assignments were made represent residues or substructures typically found in *N*- or *O*-linked glycans. Using the assigned NMR parameters as a basis, residues contained in parent structures were classified according to their residue type and glycosidic substitution sites using a relatively simple K-Nearest Neighbor pattern recognition approach. The method was able to correctly assign 99% of 77 "test residues" to their correct structural class using the full set of 19 assigned parameters as a basis. Similar correlations made between data and structure were less successful when reduced variable sets selected on the basis of SIMCA optimization were used.

INTRODUCTION

Recently we have shown that a variety of pattern recognition methods may be used to correlate structures of oligosaccharides peracetylated with ^{13}C-enriched reagents with parameters taken from one- and two-dimensional NMR experiments[1–6]. The parameters which include the coupling constant, $^{3}J_{H\text{-}1,H\text{-}2}$, the complete set of proton chemical shifts, and the resonance chemical shifts of the ^{13}C-substituted carbonyl carbons form a fingerprint characteristic of each residue within the parent structure. The pattern-recognition methods, including SIMCA class modeling, the K-Nearest Neighbor method and Principal Component analysis[7–15], utilize a comprehensive "best-fit" approach in correlating data of previously unseen structures to data of known structures already existing in a spectral library. This general, multiple-variable approach for determining complex carbohydrate structure contrasts with other NMR methods, which historically have used

Correspondence to: Professor W.J. Goux, Department of Chemistry, University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083-0688, USA.

0008-6215/93/$06.00 © 1993 – Elsevier Science Publishers B.V. All rights reserved

only the chemical shifts and coupling constants of one or two "reporter group" resonances as a means of identifying nonderivatized structures[16–22]. While for the former method structures are overdetermined by their comprehensive set of variables, structures may be underdetermined using the reporter group method, particularly if one of the reporter group resonances overlaps with resonances arising from the residual solvent or other pyranosyl ring protons. Although recently introduced two-dimensional NMR methods can be used to assign any such hidden resonances as well as providing complimentary assignments of other proton resonances, few attempts have been made to correlate the data to structure in a comprehensive manner[23–26].

In the present report we use one- and two-dimensional NMR methods to completely assign the proton and carbonyl carbon resonances of peracetylated mono- and oligo-saccharide derivatives reduced with sodium borohydride prior to peracetylation. Each of the native structures studied represent residues or substructures typically found in *N*- or *O*-linked glycans. We show that spectral assignments for these compounds are greatly facilitated owing to the absence of a mixture of anomeric ring forms at the previously reducing terminal residue. A 1-KNN method is shown to be a highly effective means of correlating new and previously assigned NMR data to classes of similar structures, each of which depends on residue type as well as sites of glycosidic substitution to neighboring residues in the parent structure. This simple method is shown to be a particularly useful way of structural determination when the size and diversity of classes of similar structure limits the applicability of statistically based methods, such as SIMCA class modeling.

EXPERIMENTAL

Materials and methods.—2-Acetamido-4-*O*-{3-*O*-(α-D-mannopyranosyl)-α-D-mannopyranosyl}-D-glucopyranose [α-Man-(1 → 3)-β-Man-(1 → 4)-GlcNAc] was purchased from BioCarb Chemicals (Lund, Sweden). (1,1′-$^{13}C_2$)Acetic anhydride was purchased from Isotec, Inc. (Miamisburg, OH). All other saccharides and reagent grade chemicals were purchased from Sigma Chemical Co. (St. Louis, MO). Reduction of the saccharides with $NaBH_4$, followed by peracetylation with ^{13}C-enriched acetic anhydride, was carried out using previously published procedures[2,4,27]. NMR samples were prepared in 5-mm sample tubes, using $CDCl_3$ as a solvent.

NMR methods.—All spectra were acquired at 11.7 T on a General Electric GN-500 NMR spectrometer. Normal COSY spectra were acquired in a 512 × 1K data array using 4 scans per t_1 experiment and a 3-s delay time between consecutive scans. COSY spectra weighted to emphasize long-range couplings (DCOSY) were similarly acquired with 16 scans per t_1 experiment and a delay Δ following both excitation pulses of 100 ms. COLOC and conventional carbon-detected carbon–proton correlation spectra were acquired in a 512 × 1K data array using 8

or 16 scans per t_1 experiment and a 3-s delay time between consecutive scans. The delay times immediately preceding and following the final observe pulse (Δ_1 and Δ_2) were lengthened to 160 and 110 ms in order to emphasize long-range couplings between labeled carbonyl carbons and pyranosyl ring protons. Hypercomplex homonuclear Hartmann–Hahn (HOHAHA) experiments were similarly carried out with 32 scans per t_1 value, a 3-s delay between scans, a 2.5-ms trim pulse and a 90-ms spin-lock mixing time[28–31].

Data representation and K-Nearest Neighbor (KNN) classification.—Each saccharide residue, either occurring as a monosaccharide or as a residue in a larger parent structure, was characterized by a set of 19 assigned parameters taken from one- and two-dimensional NMR experiments. These parameters included the assigned coupling constant $^3J_{H-1,H-2}$ and the chemical shifts for protons H-1 to H-6, the acetoxyl methyl protons (C-1–AcMe, C-2–AcMe, etc.) and carbonyl carbons (C-1–Ac, C-2–Ac, etc.) of the acetoxyl substituents (eq *1*).

$$[X] = \text{objects} \overset{\text{NMR variables}}{\begin{bmatrix} ^3J_{H-1,H-2} & \text{C-1–Ac} & \text{H-1} & \text{C-1–AcMe} & \text{C-2–Ac} & \text{H-2} & \text{C-2–AcMe} & \cdots \\ ^3J_{H-1,H-2} & \text{C-1–Ac} & \text{H-1} & \text{C-1–AcMe} & \text{C-2–Ac} & \text{H-2} & \text{C-2–AcMe} & \cdots \\ ^3J_{H-1,H-2} & \text{C-1–Ac} & \text{H-1} & \text{C-1–AcMe} & \text{C-2–Ac} & \text{H-2} & \text{C-2–AcMe} & \cdots \\ ^3J_{H-1,H-2} & \text{C-1–Ac} & \text{H-1} & \text{C-1–AcMe} & \text{C-2–Ac} & \text{H-2} & \text{C-2–AcMe} & \cdots \\ ^3J_{H-1,H-2} & \text{C-1–Ac} & \text{H-1} & \text{C-1–AcMe} & \text{C-2–Ac} & \text{H-2} & \text{C-2–AcMe} & \cdots \end{bmatrix}} \quad (1)$$

NMR shift data for a particular acetoxyl group missing as a result of having its hydroxyl in the parent compound participating in a glycosidic bond or (in the case of OH-5) in a hemiacetal bond were replaced by chemical shifts differing from the average shift for that variable by 10 standard deviations. The data set was then transformed such that each variable was mean centered and autoscaled to unit variance[7–10]. The complete data set consisted of data for the 16 residues contained in the 10 compounds studied herein in addition to data taken from previous studies, for a total of 99 residues.

Using a KNN approach, each residue in the data set was classified according to the structure of a second residue in the data set having NMR parameters which most closely resembled those of the first. The resemblance between any two parameter sets was based upon the Euclidean distance between them in space described by the total number of variables characterizing each residue[7]. Classifications made in this manner were "correct" if the KNN was of the same residue type, and had similar glycosidic linkages formed between it and neighboring residues in the parent compound. Residues having a unique structure among the other 98 residues in the initial data set were not included in the KNN calculations.

In previous work we have found that results of the KNN classification can often be improved when those variables which contribute little to distinguishing between groups of two different residue types are eliminated from the data set[1–6]. For the purpose of evaluating variable intergroup variance, the data set was first divided into separate classes, each consisting of 16 α-D-glucose, 4 α-D-galactose, 18

β-D-galactose, 20 α-D-mannose, 11 β-D-glucose, 5 2-acetamido-2-deoxy-α-D-glucose, 10 2-acetamido-2-deoxy-β-D-glucose, and 4 2-acetamido-2-deoxy-D-galactose residues. In order to maximize intraclass homogeneity, shift data missing as a result of having a hydroxyl in the parent compound participating in glycosidic or hemiacetal bonds were replaced by an average chemical shifts for that variable. The classes were modeled independently using the SIMCA algorithm, according to the expression[7,8,10–13]

$$x_{ik}^{(q)} = \alpha_k^{(q)} + \sum_{a=1}^{A_q} \beta_{ak}^{(q)} \theta_{ki}^{(q)} + \epsilon_{ik}^{(q)} \qquad (2)$$

where $x_{ik}^{(q)}$ contains elements of the autoscaled original data matrix for class q, $\alpha_k^{(q)}$ contains means with respect to variable k, $\beta_{ak}^{(q)}$ contains the loadings of the A_q principal components, $\theta_{ki}^{(q)}$ contains the coordinates of the transformed points (scores), and $\epsilon_{ik}^{(q)}$ contain residuals or differences between the actual components of the data matrix and the sum of the first two terms on the right. Each class of residues was modeled using two principal components. According to methods outlined in previous publications, various data sets were constructed, keeping those variables having a large variable discriminatory power, $DP^{(r,q)}(k)$, for distinguishing between classes, r and q[3–6]. Since discriminatory power is defined in terms of pairwise interaction between classes, the average of pairwise interactions was used as a criterion for the selection of variables. While some of the reduced variable sets were formed keeping only those variables with the largest average discriminatory power, others were formed using variable discriminatory power as only a guiding criterion. In the latter case variables having lower discriminatory powers were substituted into reduced variable set KNN calculations based on the types of misassignments made in a previous iteration.

Principal component (PC) analysis was carried out on each of the reduced data sets. Finally PC plots were constructed by plotting the scores of the data using as axes the two largest principal components (representing 62% of the data variance)[7]. All computations were carried out on an IBM PC-compatible microcomputer using the SIMCA 3B software package obtained from Principal Data Components (Columbia, MO). Computational routines needed to evaluate variable discriminatory power were written in BASIC.

RESULTS AND DISCUSSION

Assignment of resonances in reduced, peracetylated carbohydrate derivatives.—The general strategy used in assigning the resonances of the carbohydrate backbone protons, the acetyl methyl protons, and the ^{13}C-substituted carbonyl carbons has already been described for oligosaccharide derivatives not reduced prior to peracetylation with (1,1′-$^{13}C_2$)acetic anhydride[1–6]. Using this assignment strategy, resonances arising from the carbohydrate backbone are first assigned using a

combination of homonuclear correlation spectroscopies, including COSY, COSY optimized for the observation of long-range couplings (DCOSY), and HOHAHA[28,29]. With few exceptions these assignments begin with the initial assignment of the anomeric proton resonances. Assignments of these resonances are facilitated by their unique chemical shift and doublet fine structure arising from coupling between H-1 and H-2. Resonances arising from the remaining protons of the carbohydrate backbone may then be assigned by tracing out the coupling network observed in the COSY and DCOSY spectra. If needed, the HOHAHA experiment provides through space correlations across several bonds in the same residue, providing a check on the assignments made via three-bond couplings using the COSY and DCOSY experiments[30,31]. Once the proton spectrum of a peracetylated carbohydrate has been completely assigned, the ^{13}C-substituted carbonyl carbon resonances can be assigned using heteronuclear correlation spectroscopy. Finally the assignment of the acetyl methyl proton resonances is made via their correlation to the assigned carbonyl carbon resonances.

A standard technique used in the structural determination of oligosaccharides using gas–liquid chromatography (GLC) or GLC–mass spectrometry is borohydride reduction of the terminal reducing residue prior to peracetylation or permethylation[27]. This procedure eliminates the possible formation of a mixture of anomeric forms, resulting in a simpler interpretation of the chromatographic data. Likewise, one would expect the interpretation of the carbonyl region of the ^{13}C NMR spectrum of peracetylated oligosaccharide derivatives reduced with borohydride prior to derivatization to be similarly simplified, such that only one signal is observed for each acetylation site. At the same time, assignment of resonances of backbone protons of the reduced derivatized residue becomes more difficult without the unique shift and coupling fine structure characteristic of the anomeric proton resonance. The COSY spectrum of Fig. 1 illustrates how other resonances having unique shifts can be used for assigning backbone proton resonances of the previously reducing glucose residue of reduced peracetylated lactose. Based on resonance assignments made previously for peracetylated α-lactose, the two H-6 resonances and the aglycone H-4 resonance of the glucose moiety have shifts upfield of about 4.5 ppm. In the COSY spectrum of the reduced, peracetylated derivative only the resonance at 5.03 ppm, assignable to H-5, correlates with three upfield resonances at 4.05, 4.08, and 4.44 ppm. The off-diagonal contour connecting the 4.05 and the 4.44 resonances is consistent with the expected coupling between the two hydroxymethyl H-6 protons, leaving the resonance at 4.08 ppm assignable to H-4. Once the H-4 resonance has been assigned, the H-3, H-2, and the H-1 resonances may be identified in a straightforward manner. Resonance assignments were made in the previously nonreducing galactose moiety by initially assigning the doublet at 4.65 ppm to the anomeric proton resonance. This assignment, as well as those of the other galactose backbone protons that follow from the COSY spectrum, are in reasonable agreement with those made for previously studied peracetylated galactose residues of similar structure[2,6].

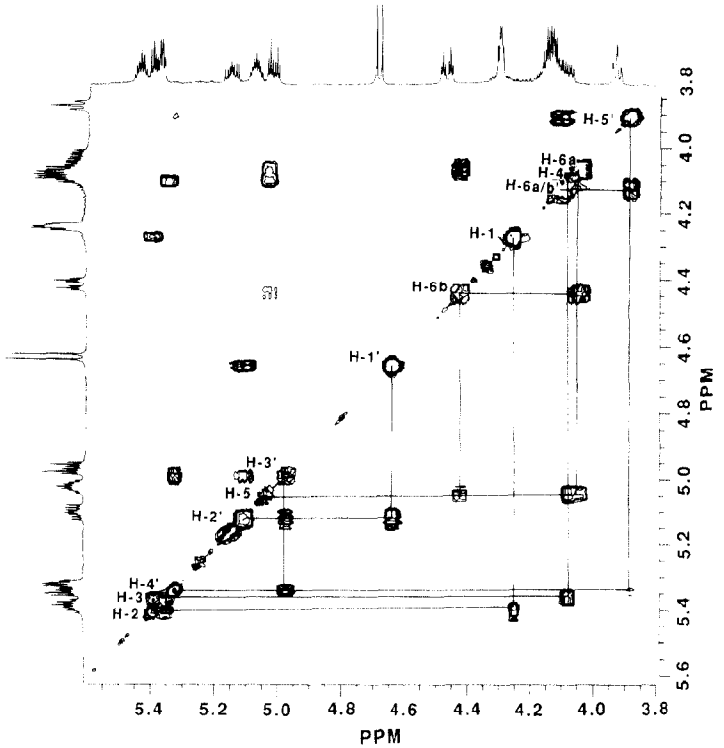

Fig. 1. The two-dimensional COSY spectrum of lactose, reduced with borohydride prior to peracetylation with (1,1'-$^{13}C_2$)acetic anhydride (**8**). Lines in the figure demonstrate the method used in tracing out the coupling network for each residue in the parent structure.

Once the proton spectrum of a peracetylated carbohydrate has been completely assigned, the ^{13}C-substituted carbonyl carbons can be assigned using heteronuclear correlation spectroscopy. Fig. 2 demonstrates how the COLOC experiment[29] can be used to correlate previously assigned pyranosyl ring proton resonances and heretofore unassigned acetoxyl methyl proton resonances to nearest-neighbor carbonyl carbon resonances. Assignment data for the carbonyl carbon, the acetoxyl methyl proton, and the carbohydrate backbone proton resonances are summarized in Table I for all borohydride reduced peracetylated compounds studied.

Principal component representation of NMR data.—The ultimate goal of this and previous studies has been to establish a library of NMR parameters of various peracetylated carbohydrate structures and substructures from which unknown structures may be identified. Accordingly, each residue of a parent structure is represented by a vector of parameters whose components include $^{3}J_{H\text{-}1,H\text{-}2}$ and the chemical shifts of all of the proton resonances and the ^{13}C-substituted carbon carbonyl resonances taken from conventional one- and two-dimensional NMR experiments (eq *1*). Data for the 16 residues contained in the 10 compounds of Table I and for the 83 residues reported in previous studies are shown in Fig. 3A

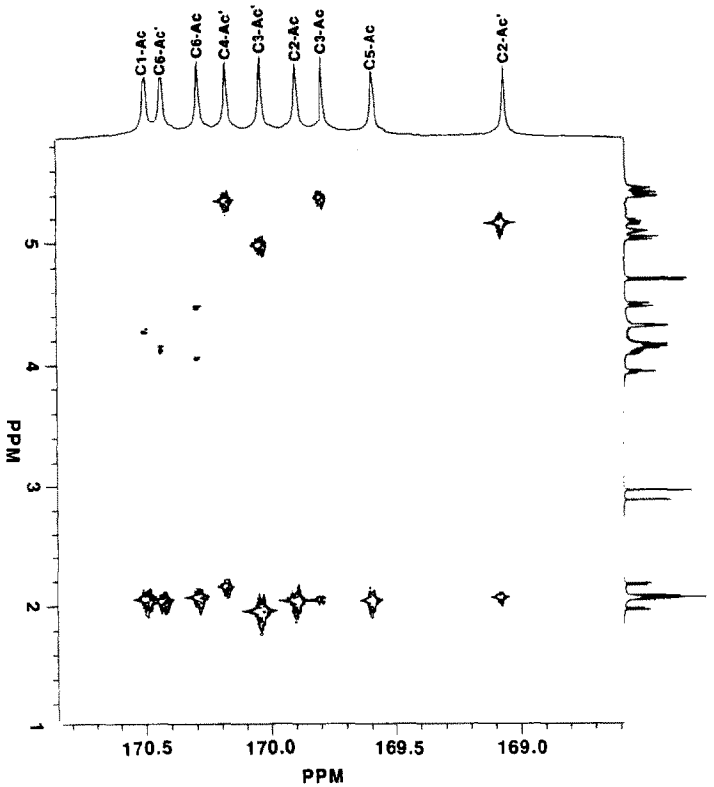

Fig. 2. The two-dimensional COLOC spectrum of **8**. Normal ^{13}C and ^{1}H NMR spectra are shown along the horizontal and vertical axes. Notations for carbonyl carbons is similar to those used in Table I.

in the form of a principal component plot[7]. The plot is a two-dimensional representation of the data, where the absissa and ordinate represent those linear combinations of original variables which point in the directions of greatest variance of the data. Residues in the plot appear reasonably well segregated into groups where each of the members of a group is structurally similar with respect to sites of glycosidic bond substitution. This segregation arises as a result of the unusual shifts which were substituted for missing acetoxyl resonances in the original data set (see Experimental). Similarly, peracetylated residues and reduced peracetylated residues are seen to lie in separate regions of the plot as a result of the unusual C-5–Ac and C-5–AcMe shifts given to those residues whose C-5 hydroxyl is involved in pyranose ring formation. Within each of the groups representing residues with similar sites of acetoxyl substitution there is visible segregation into subgroups of different residue types. This is best shown in the principal component plot of the peracetylated, nonreducing terminal residues (Fig. 3B).

Correlation of residue structure and NMR data using pattern recognition methods. —In previous reports we have shown both the SIMCA and KNN pattern recognition methods can be used to correlate the NMR data of unknown samples to those

TABLE I

Summary of NMR chemical shift and coupling constant data on reduced, peracetylated carbohydrate derivatives [a]

Acetoxyl substituent on	Reduced peracetylated derivative							
	1 Glc(R)				**2** Gal(R)			
	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$
C-1	170.32	4.27, 3.96	1.98	3.15	170.34	4.24, 3.78	1.97	3.15
C-2	169.94	5.17	2.01		170.18	5.25	2.03	
C-3	169.70	5.36	1.99		169.67	5.30	2.05	
C-4	169.74	5.35	2.06		169.67	5.30	2.05	
C-5	169.66	4.97	1.98		170.18	5.25	2.03	
C-6	170.42	4.16, 4.06	2.00		170.34	4.24, 3.78	1.97	
	3 Man(R)				**4** GalNAc(R)			
	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$
C-1	170.49	4.15, 4.03	2.00	3.16	170.50	3.97, 3.87	1.98	3.14
C-2	169.82	5.02	2.02			4.54		
C-3	169.61	5.39	2.05		169.26	5.29	2.03	
C-4	169.61	5.39	2.05		169.69	5.18	2.09	
C-5	169.82	5.02	2.02		170.20	5.21	2.03	
C-6	170.49	4.15, 4.03	2.00		170.39	4.25, 3.76	1.96	
	5 GlcNAc(R)				**6** α-Glc-(1 → 4)-Glc(R)			
	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$
C-1	170.49	4.03	2.00	3.15	170.15	4.25, 3.90	1.99	3.15
C-2		4.47			169.85	5.34	2.06	
C-3	170.43	5.24	2.03		169.59	5.36	2.05	
C-4	169.96	5.35	2.05			4.06		
C-5	169.64	5.08	1.99		169.47	5.12	2.04	
C-6	170.45	4.19, 4.11	2.00		170.19	4.48, 4.20	2.05	
C-1′						5.21		3.71
C-2′					170.24	4.87	2.05	
C-3′					169.63	5.32	1.95	
C-4′					169.33	5.02	2.00	
C-5′						4.18		
C-6′					170.38	4.24, 4.00	2.04	
	7 α-Glc-(1 → 6)-Glc(R)				**8** β-Gal-(1 → 4)-Glc(R)			
	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$
C-1	170.50	4.32, 3.98	2.01	3.15	170.35	4.26	2.05	3.15
C-2	170.14	5.30	2.07		169.77	5.39	2.05	
C-3	169.75	5.40	2.02		169.68	5.36	2.04	
C-4	169.80	5.40	2.09			4.08		
C-5	169.49	5.01	1.99		169.51	5.03	2.03	
C-6		3.67, 3.64			170.18	4.44, 4.05	2.06	
C-1′		5.03		3.60		4.65		7.84
C-2′	170.33	4.87	2.05		169.01	5.11	2.06	
C-3′	169.84	5.38	1.94		169.90	4.99	1.96	
C-4′	169.57	5.02	1.98		170.08	5.33	2.15	
C-5′		3.98				3.89		
C-6′	170.61	4.24, 4.07	2.09		170.30	4.12	2.03	

TABLE I (continued)

Acetoxyl substituent on	Reduced peracetylated derivative							
	9 β-Gal-(1 → 4)-GlcNAc(R)				**10** α-Man-(1 → 3)-β-Man-(1 → 4)-GlcNAc(R)			
	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$	δ^{13}C-Ac	δ^{1}H-PR	δ^{1}H-AcMe	$J_{H\text{-}1,H\text{-}2}$
C-1	170.58	4.31, 3.99	2.04	3.15	170.66	4.31, 3.99	2.01	3.15
C-2		4.61				4.55		
C-3	170.57	5.22	2.06		170.45	5.13	2.00	
C-4		4.11				4.12		
C-5	169.61	5.06	2.02		169.78	5.15	1.99	
C-6	170.28	4.46, 4.05	2.06		170.40	4.47, 4.05	2.00	
C-1′		4.63		7.95		4.78		3.27
C-2′	169.14	5.14	2.07		170.62	5.43	2.20	
C-3′	170.01	4.99	1.95			3.89		
C-4′	170.17	5.35	2.15		169.73	5.17	2.08	
C-5′		3.88				3.53		
C-6′	170.45	4.14	2.04		170.77	4.24, 4.14	2.03	
C-1‴						4.96		3.30
C-2‴					170.00	4.96	2.07	
C-3‴					169.50	5.15	1.93	
C-4‴					169.80	5.28	1.97	
C-5‴						4.07		
C-6‴					170.61	4.26, 4.16	2.03	

[a] All chemical shifts are reported downfield with respect to Me_4Si used as the internal standard. The abbreviations used are ^{1}H-PR for pyranosyl ring proton or the methoxy protons at C-1 or C-6; ^{1}H-AcMe for acetoxyl methyl proton: ^{13}C-Ac for acetoxyl carbonyl carbon.

of known structures contained in a spectral library[2–6]. Both of these methods provide a comprehensive best fit of unknown data to known data in contrast to many other methods which rely on the fitting of a few experimental parameters within a specified tolerance[22]. Using the SIMCA method, the data from each group of residues of similar structure are independently modeled according to their principal components. Based on a set of statistical criteria, it is then determined to which model data from residues of unknown structure best fit. Because of its statistical nature, the SIMCA procedure works best when the diversity in the structures and associated data variance within each group matches the diversity of unknown structures being identified. Such will be the case when each group of structurally similar compounds is composed of relatively large numbers of residues coming from a variety of parent compounds. In previous work, we found the SIMCA method can quite successfully be used to classify a set of test residues according to their residue type, irrespective of glycosidic substitution sites on neighboring residues. Each of the groups into which test residues were classified was large and structurally diverse enough to allow for the formation of a reliable model. In the present study these large groups have now been divided into much smaller groups, where each group takes into account residue type, glycosidic substitution sites, and reduction prior to peracetylation. As a result there are

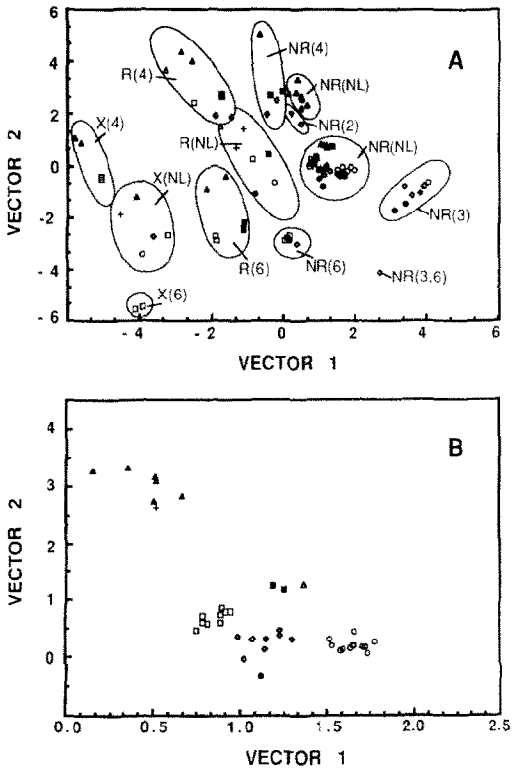

Fig. 3. (A) Principal component plot of the 99 residue data set. Symbols designate different residues types according to whether they are α-D-glucose (□), β-D-glucose (■), α-D-galactose (●), β-D-galactose (○), α-D-mannose (◇), 2-acetamido-2-deoxy-α-D-glucose (△), 2-acetamido-2-deoxy-β-D-glucose (▲), or 2-acetamido-2-deoxy-α- or β-D-galactose (+). Residues appear grouped according to whether they occur as terminal reducing (R), nonreducing (NR) or borohydride reduced residues (X) prior to peracetylation. Reduced residues are denoted symbolically according to their residue type only. Residues which are glycosidically linked are denoted by placing their substituted hydroxyl in parenthesis [i.e., NR(3) residues are nonreducing and have a glycosidic linkage through OH-3 in the native structure]. (B) PC Plot showing nonreducing residues having no glycosidic substitutions [NR(NL)].

insufficient numbers of member residues within each group to allow the SIMCA method to be used as a method of residue classification. An alternative means of assigning structures to a set of NMR data is the KNN method. Structures are assigned based on the similarity of descriptive NMR data vectors to those of at least one residue already existing in the data set. The degree to which unknown data resembles known data is measured by the Euclidean distance in NMR parameter space. Since structural assignments may be made on a pairwise basis, there is no need for the formation of larger classes of structurally similar residues, as was the case using the SIMCA method.

A measure of the reliability of the KNN model can be made by assigning each member of the existing data set a structure based on its similarity with another member. Assignments are considered "correct" when the test residue is found to

most resemble another member residue of the data set of the same residue type and forming similar glycosidic linkages to neighboring residues. Residues not having a structurally similar partner are eliminated as test residues. Results of a 1-KNN calculation show that only one residue of the 77 test residues is assigned to an incorrect structure when each of the residues is represented by its full compliment of 19 NMR variables. The single misassignment occurred when data representing a peracetylated α-galactosyl residue glycosidically linked to a neighboring residue at its 3-pyranose ring position was incorrectly assigned a structure corresponding to a similarly glycosidically linked α-mannoside.

In previous studies it was shown that the effectiveness of the KNN method can often be improved by eliminating those NMR variables from the vector representation of a residue which have little value in discriminating between structures belonging to different classes[3–6]. Because of the nature of the misassignment when the full compliment of variables was used, it seems most reasonable to assume that the results of a KNN calculation would be improved by selecting those variables which emphasize differences between different residue types, with less emphasis on distinguishing between structural classes having different sites of glycosidic substitution. As was the case in previous studies, variables included in variable subsets were selected on the basis of their average discriminatory power between classes of identical residue types[6]. These variable subsets along with the results of the 1-KNN calculations using them are summarized in Table II. The data show that all of the test residues are correctly assigned in a KNN calculation carried out with a variable subset formed by eliminating seven of the least discriminatory variables (set DP-12). However, when three additional variables having low discriminatory power are eliminated from the variable subset, the number of misassignments made in the KNN calculation significantly increased. The majority of these misassignments occurred when data representing 6-*O*-glycosidically linked test residues were misassigned to non-glycosidically linked structures or when data representing residues having a pyranose ring structure were misassigned to re-

TABLE II

Results of K-Nearest Neighbor calculations

Data set	Variables	Number misassigned (% correctly classified)
VAR-19	Complete data set	1 (99)
DP-12	$J_{H\text{-}1,H\text{-}2}$, H-1, H-2, H-3, H-4, H-5, H-6 C-1–Ac, C-1–AcMe, C-2–AcMe, C-3–Ac, C-4–Ac	0(100)
DP-9	$J_{H\text{-}1,H\text{-}2}$, H-1, H-2, H-3, H-4 C-1–AcMe, C-2–AcMe, C-3–Ac, C-4–Ac	8 (90)
SP-6/1	$J_{H\text{-}1,H\text{-}2}$, H-2, H-3, H-4, H-5, H-6	0(100)
SP-6/4	$J_{H\text{-}1,H\text{-}2}$, H-2, H-3, H-4, C-5–Ac, C-6–Ac	1 (99)
SP-6/5	$J_{H\text{-}1,H\text{-}2}$, H-2, H-3, H-4, C-5–AcMe, C-6–AcMe	3 (96)

duced linear structures. The nature or these errors suggests that at least some of the variables associated with the C-5 and C-6 sites are important in correlating the NMR data to structure.

Further insight as to which variables are important in correlating the NMR data with structure can be gained by using variable discriminatory power as a guiding rather than an absolute criterion in the selection of variable subsets used in carrying out classifications by the KNN method. Using a trial-and-error approach we found that successful results may be achieved by including, in the variable subsets, the coupling constant, $^{3}J_{\text{H-1,H-2}}$, and one of the chemical shift parameters associated with structural sites C-2 to C-6 (subsets SP-6/1, SP-6/4 and SP-6/5 of Table II). Significantly poorer results were obtained when the variable subset consisted of fewer than six variables or when the coupling constant $^{3}J_{\text{H-1,H-2}}$ was replaced by the chemical shift associated with the H-1 resonance. This result is rather surprising in light of the fact that $^{3}J_{\text{H-1,H-2}}$ and the chemical shifts associated with H-1 and H-2 are used almost exclusively in determining structures from ^{1}H spectra of nonderivatized carbohydrates in aqueous solvents.

CONCLUSIONS

In previous reports we have demonstrated the feasibility of using the full complement of NMR parameters assigned from one- and two-dimensional NMR experiments to determine structures of peracetylated oligosaccharides[1–6]. In comparison to carrying out the same task from spectra of nonderivatized samples in aqueous solvents, the assignment of resonances in spectra of peracetylated oligosaccharides is relatively easy. This results from the absence of a solvent resonance in the ^{1}H NMR spectra of these derivatized compounds in deuterated organic solvents and because derivitization tends to "spread out" the ^{1}H resonances, facilitating the assignment task. In the present report we have shown that borohydride reduction prior to peracetylation further simplifies the assignment task by eliminating the possibility of having to carry out resonance assignments on an anomeric mixture of compounds. We have also shown that carbohydrate residue structures may be correlated with the NMR data using a rather simple 1-KNN pattern recognition approach. The advantage of this method over the SIMCA method used in previous studies arises from the fact that larger classes of structurally similar residues need not be formed from a library of data having a limiting number of residues. As we have noted previously, improved results may be obtained from the KNN method using variable subsets where variables are selected using average variable discriminatory power as a guide for variable selection. In order for the method to be reasonably successful ($>95\%$ accurate) $^{3}J_{\text{H-1,H-2}}$ and chemical shift parameters associated with structural sites C-1 to C-6 must be retained in the variable subset. The method was able to classify test data with 100% reliability when only $^{3}J_{\text{H-1,H-2}}$ and chemical shifts of ^{1}H resonances associated with H-2 to H-6 were included in the variable subset. This result

suggests that, unless correlation of proton resonances to those of ^{13}C-substituted carbonyl carbons is useful as an aid in making proton assignments, the assignment of carbonyl carbon resonances may be unnecessary for structural determination.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the Robert A. Welch Foundation (AT-1162).

REFERENCES

1 W.J. Goux and C.J. Unkefer, *Carbohydr. Res.*, 159 (1987) 191–210.
2 W.J. Goux, *Carbohydr. Res.*, 184 (1988) 47–65.
3 W.J. Goux, *J. Magn. Res.*, 85 (1990) 457–469.
4 G. Okide, D.S. Weber, and W.J. Goux, *J. Magn. Res.*, 96 (1992) 526–540.
5 W.J. Goux, in J.W. Finley, S.J. Schmidt, and A.S. Serrianni (Eds.), *NMR Applications in Biopolymers (Basic Life Sciences)*, Plenum Press, New York, 1990, Vol. 56, pp 47–62.
6 D.S. Weber and W.J. Goux, *Carbohydr. Res.*, 233 (1992) 65–80.
7 M.A. Sharaf, D.L. Illman, and B.R. Kowalski, *Chemometrics (Chemical Analysis)*, Wiley, New York, 1986, Vol. 82, pp 179–296.
8 C. Albano, G. Blomquist, W. Dunn, W., III, U. Edlund, B. Eliasson, E. Johansson, B. Norden, M. Sjostrom, B. Soderstrom, and S. Wold, in A. Vermavwori (Ed.), *27th Intl. Congr. Pure Appl. Chem.*, Pergamon Press, New York, 1979, pp 377–386.
9 B.R. Kowalski, *Anal. Chem.*, 47 (1975) 1152A–1162A.
10 S. Wold and M. Sjostrom, *ACS Symp. Ser.*, 52 (1976) 243–252.
11 M. Sjostrom and U. Edlund, *J. Magn. Reson.*, 25 (1977) 285–297.
12 U. Edlund and S. Wold, *J. Magn. Reson.*, 37 (1980) 183–194.
13 S. Wold, *Pattern Recog.*, 8 (1976) 127–139.
14 B.R. Kowalsky and C.F. Bender, *Anal. Chem.*, 44 (1972) 1405–1411.
15 P.C. Jurs, *Science*, 232 (1986) 1219–1224.
16 J. Montreuil, *Adv. Carbohydr. Chem. Biochem.*, 37 (1980) 157–223.
17 D.A. Cumming, R.N. Shah, J.J. Krepinsky, A.A. Grey, and J.P. Carver, *Biochemistry*, 26 (1987) 6655–6676.
18 J.F.G. Vliegenthart, H. van Halbeek, and L. Dorland, *Pure Appl. Chem.*, 53 (1981) 45–77.
19 J.F.G. Vliegenthart, L. Dorland, and H. van Halbeek, *Adv. Carbohydr. Chem. Biochem.*, 41 (1983) 209–374.
20 D.R. Anderson and W.J. Grimes, *Anal. Biochem.*, 146 (1985) 13–22.
21 E.F. Hounsell, D.J. Wright, A.S.R. Donald, and J. Feeney, *Biochem. J.*, 223 (1984) 129–143.
22 E.F. Hounsell and D.J. Wright, *Carbohydr. Res.*, 205 (1990) 19–29.
23 J. Dabrowski, U. Dabrowski, P. Hanfland, M. Kordowicz, and W.E. Hull, *Magn. Reson. Chem.*, 24 (1986) 59–69.
24 E. Berman, U. Dabrowski, and J. Dabrowski, *Carbohydr. Res.* 176 (1988) 1–15.
25 M. Ikura and K. Hikichi, *Carbohydr. Res.*, 163 (1987) 1–8.
26 S.W. Homans, R.A. Dwek, J. Boyd, N. Soffe, and T.W. Rademacher, *Proc. Natl. Acad. Sci. USA*, 84 (1987) 1202–1205.
27 W.S. York, A.G. Darvill, M. McNeil, T.T. Stevenson, and P. Albersheim, *Methods Enzymol.*, 118 (1985) 3–40.
28 A. Bax, *Two-Dimensional Nuclear Magnetic Resonance in Liquids*, Delft University Press, Delft, Holland, 1982, pp 50–98.
29 G.E. Martin and A.S. Zektzer, *Two-Dimensional NMR Methods for Establishing Molecular Connectivity*, VCH Publishers, New York, 1988, pp 58–347.
30 R.A. Byrd, W. Egan, M.F. Summers, and A. Bax, *Carbohydr. Res.*, 166 (1987) 47–58.
31 L. Lerner and A. Bax, *Carbohydr. Res.*, 166 (1987) 35–46.