Carbohydrate Research, 233 (1992) 65–80
Elsevier Science Publishers B.V., Amsterdam

The application of NMR-pattern-recognition methods to the classification of peracetylated oligosaccharide residues: effects of intraclass structure

Denise S. Weber and Warren J. Goux

Department of Chemistry, University of Texas at Dallas, P.O. Box 830688, Richardson, Texas 75083-0688 (USA)

(Received November 25th, 1991; accepted February 25th, 1992)

ABSTRACT

In the present study a variety of homo- and hetero-nuclear correlation spectroscopies have been used to assign the acetoxyl carbonyl carbon, the pyranosyl proton and the acetoxyl methyl proton resonances of thirteen oligosaccharide derivatives peracetylated with [1,1′-^{13}C]acetic anhydride. The nonderivatized forms of these structures occur as D-glucose, 2-acetamido-2-deoxy-D-glucose, D-galactose and 2-acetamido-2-deoxy-D-galactose containing substructures of *O*-linked glycans. On the basis of the assigned NMR variables, two pattern recognition methods, K-nearest neighbor (KNN) and SIMCA, were used to classify residues contained in these and previously studied peracetylated derivatives according to their structure and anomeric ring configuration. It was found that the SIMCA method was able to classify residues into one of eight structurally homogeneous classes with greater than 99% accuracy. In contrast, the KNN method proved to be most successful in classifying residues into one of six larger more structurally diverse classes, where some of the classes were formed by members of the same residue type but having different anomeric ring configurations. While the performance of the KNN method was improved by using variable subsets as a basis for classification, SIMCA performed best using the full compliment of 15 NMR variables. Neither of the methods was able to classify residues well when only proton chemical shifts and coupling constants were used to assign structures. This suggests that those previous methods which have traditionally used limited ^{1}H NMR data to make structural assignments of carbohydrate residues may be significantly improved by using complimentary ^{13}C NMR data.

INTRODUCTION

In the recent past we have tried to establish a correlation between the structures of peracetylated carbohydrate derivatives, which are ^{13}C-substituted at their carbonyl carbons, and a comprehensive set of NMR parameters taken from homo- and hetero-nuclear 2D NMR experiments[1–5]. The parameters, which form a

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

fingerprint characteristic of each residue within the parent structure, include the coupling constant, $J_{H\text{-}1,H\text{-}2}$, and the chemical shifts of the backbone protons of the carbohydrate, the acetoxyl methyl protons, and the ^{13}C-substituted carbonyl carbons. Once a spectral library representative of the different types of substructures has been established, residues in a peracetylated oligosaccharide of unknown structure can be identified based on the similarity of their spectral parameters with those of other residues existing in the data set. A decision as to which residue or group of residues an unknown residue most resembles can be facilitated using a variety of pattern-recognition techniques including the K-nearest neighbor (KNN) method, principal component analysis, or SIMCA class modeling[6–14]. Parameters implicit in the class modeling method allow for the elimination of those NMR variables which are of lesser value in discriminating between classes of different residue types, thus impoving the ability with which an unknown residue can be correctly classified. This general multiple variable approach for determining complex carbohydrate structure contrasts with other NMR methods which historically have used only the chemical shifts and coupling constants of one or two "reporter-group" protons as a means of residue identification[15–21]. Because the latter method utilizes one-dimensional spectra of underivatized oligosaccharides in a deuterated aqueous solvent, possible complications may arise if the reporter-group resonances overlap with those of the residual solvent or other resonance in the molecule. More elegant two-dimensional experiments have allowed for the assignment of any such hidden resonances as well as provided complimentary assignments of backbone protons in the same and neighboring residues[22–25]. To date no attempts have been made to correlate residue structures with the complimentary parameters derived from these experiments[21]. Recently a more comprehensive form of pattern recognition has emerged which attempts to solve the problems associated with resonance degeneracies that occur in one-dimensional spectra of nonderivitzed carbohydrates in aqueous solvents. Rather than a few recognizable spectral features being used as a means of structural identification, entire proton spectra of known compounds are digitized and presented to a neural network[26]. Repeated presentation of the data allows the network to optimize its internal parameters such that ultimately it can correlate the appearance of an entire spectrum with one of the compounds presented during the training session. An appropriately optimized model may then be used to classify unknown spectra, should they match those already contained in the library. In essence, multiple features contained in the spectrum of one compounds are weighted based on the limited data contained in other spectra presented during the same learning session. Limited flexibility is allowed for the extrapolation from the spectrum of an unknown compound to another compound of similar structure not contained in the original spectral library.

In the present study a variety of homo- and hetero-nuclear correlation spectroscopies have been used to assign the proton and carbonyl carbon resonances of peracetylated oligosaccharide derivatives whose native structures occur as sub-

structures of *O*-linked glycans. Two pattern-recognition methods, KNN and SIMCA, are then compared with respect to their ability to correctly classify according their overall structure. The comparison has been carried out both under conditions in which the total data set was divided into homogeneous classes and under conditions in which some of the classes are combined to give larger classes, each having a more diversified membership. We find that the overall success of each of the methods in correctly assigning residues to their proper structural class depends on the basis set of NMR variables used in making the classifications and on the number of residues forming each of the classes.

EXPERIMENTAL

Materials and methods.—All saccharides and reagent chemicals were purchased from Sigma Chemical Co. (St. Louis, MO). [1,1′-^{13}C]Acetic anhydride was purchased from Isotec, Inc. (Miamisburg, OH). Peracetylated carbohydrate derivatives were prepared by acetylation with [1,1′-^{13}C]acetic anhydride according to previously published methods[2]. The final products following peracetylation were methyl 2,4,6-tri-*O*-acetyl-3-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-β-D-galactopyranoside (**1**), methyl 2,4,6-tri-*O*-acetyl-3-*O*-(2,3,4,6-tetra-*O*-acetyl-α-D-galactopyranosyl)-α-D-galactopyranoside (**2**), 1,2,3,6-tetra-*O*-acetyl-4-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-α-D-mannopyranose (**3**), 1,2,3,6-tetra-*O*-acetyl-4-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-β-D-mannopyranose (**4**) methyl 2,3,6-tri-*O*-acetyl-4-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-β-D-glucopyranoside (**5**), 2-acetamido-1,4,6-tri-*O*-acetyl-2-deoxy-3-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-α-D-galactopyranose (**6**), 2-acetamido-1,3,4-tri-*O*-acetyl-2-deoxy-6-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-α-D-glucopyranose (**7**), 2-acetamido-1,3,4-tri-*O*-acetyl-2-deoxy-6-*O*-(2,3,4,6-tetra-*O*-acetyl-β-D-galactopyranosyl)-β-D-glucopyranose (**8**), methyl 2,4,6-tri-*O*-acetyl-3-*O*-(2-acetamido-3,4,6-tri-*O*-acetyl-2-deoxy-β-D-galactopyranosyl)-α-D-galactopyranoside (**9**), methyl 2,4,6-tri-*O*-acetyl-3-*O*-(2-acetamido-3,4,6-tri-*O*-acetyl-2-deoxy-β-D-glucopyranosyl)-β-D-galactopyranoside (**10**), 2-acetamido-1,3,4,6-tetra-*O*-acetyl-2-deoxy-α-D-galactopyranose (**11**), 2-acetamido-1,3,4,6-tetra-*O*-acetyl-2-deoxy-β-D-galactopyranose (**12**), and methyl 2,3,6-tri-*O*-acetyl-4-*O*-{2,4,6-tri-*O*-acetyl-3-*O*-[2-acetamido-3,6-di-*O*-acetyl-2-deoxy-4-*O*-(2,3,4,6,tetra-*O*-acetyl-β-D-galacopyranosyl)-β-D-glucopyranosyl)-β-D-galactopyranosyl)-β-D-glucopyranoside (**13**). NMR samples were prepared in 5-mm sample tubes, using $CDCl_3$ as solvent.

NMR methods.—All spectra were acquired on 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 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 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[27–30].

Two-dimensional representations of NMR data.—Each saccharide residue, either occurring as a monosaccharide or as a residue in a larger parent structure, was characterized by its set of assigned NMR parameters arranged in a 15-dimensional vector[3–5].

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

These parameters include the assigned chemical shifts for pyranosyl ring protons H-1–H-6, the acetoxyl methyl protons (C-2 AcMe, C-3 AcMe, etc.), and carbonyl carbons of substituents at C-2, C-3, C-4 and C-6 (i.e., C-2 Ac). NMR shift data for a particular acetoxyl group missing as a result of aglycon substitution were replaced with the average chemical shifts for those particular vector components. Data for those acetoxyl groups missing as a result of acetamido substitution were replaced with unique shift values unlike those for any other component ($\delta = 1.0$ ppm). Each variable was mean-centered and autoscaled to unit variance. Various data sets were then constructed within each of which those variables which did not have great interclass variance were eliminated (see below). Principal component (PC) analysis[6] 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.

Selection of variables.—The complete data set consisted of data for the twenty-four residues contained in the thirteen compounds studied herein in addition to data determined in previous studies[1–4], for a total of 80 residues. These residues were classified using two different methods. By the first of these methods (Scheme I), the complete data set was divided into eight classes comprised of 12 α-D-glucose residues, 11 β-D-glucose residues, 4 α-D-galactose residues, 16 β-D-galactose residues, 17 α-D-mannose residues, 5 2-acetamido-2-deoxy-α-D-glucose residues, 10 2-acetamido-2-deoxy-β-D-glucosamine residues and 4 2-acetamido-2-deoxy-α, β-D-galactose residues. By the second method (Scheme II), six classes were formed from 12 α-D-glucose residues, 20 α,β-D-galactose residues, 18 α,β-D-mannose residues, 11 β-D-glucose residues, 15 2-acetamido-2-deoxy-α,β-D-glucose residues

and 4 2-acetamido-2-deoxy-α,β-D-galactose residues. The classes formed under both classification schemes were modeled independently using the SIMCA algorithm, according to the expression[6,7,9–12]

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

where the $x_{ik}^{(q)}$'s are elements of the autoscaled original data matrix for class q, the $\alpha_k^{(q)}$'s are means with respect to variable , the $\beta_{ak}^{(q)}$ are the loadings of the A_q principal components, the $\theta_{ki}^{(q)}$'s are the coordinates of the transformed points (scores) and the $\epsilon_{ik}^{(q)}$'s are 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 one principal component. Previous studies have shown that variables having greater interclass variances are more important for the classification of unknown residues than are those having greater intraclass variances. Accordingly, 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, where $Dp^{(r,q)}(k)$ is defined as

$$Dp^{(r,q)}(k) = \left[\frac{S_{(r)}^{(q)}(k)^2 + S_{(q)}^{(r)}(k)^2}{S_{(q)}^{(q)}(k)^2 + S_{(r)}^{(r)}(k)^2} \right]^{1/2} - 1 \qquad (3)$$

and the residual standard deviation of a variable, k, between classes q and r is given by

$$S_{(r)}^{(q)}(k) - \left[\sum_{p=1}^{N_q} \frac{NV\, e_{p,k}^{(q)2}}{(NV - A_q) N_r} \right]^{1/2} \qquad (4)$$

Here N_r is the number of residues in class r and NV is the total number of variables defining the data. When class r is equal to class q, the residuals, $e_{p,k}^{(q)}$, are equivalent to those in eq 2. When classes r and q differ, the residuals are those obtained by least-squares fitting, using the object scores, $\theta_{i,k}$, as adjustable parameters. 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.

K-nearest neighbor (KNN) classifications.—K-nearest neighbor calculations were performed using reduced variable subsets selected on the basis of their discriminatory power[3–6,13,14]. Residues were classified according to the two classification methods using as a criterion for classification their Euclidean distance to their nearest neighbor or their two nearest neighbors.

SIMCA classifications.—Classes of residues formed according to the two classification methods were randomly divided into subsets of test objects and those objects used in the final modeling of each class. In all, twelve or fifteen of the 80 residues served as test objects when classification was carried out according to

Schemes I or II, respectively (15 or 18% of the data set). Following elimination of test objects, each of the classes was modeled independently according to eq 2 using reduced variable sets, where variables were selected on the basis of their average discriminatory power. Objects were classified on the basis of an F test at the 90% confidence level with $(NV\text{-}A_q)$ and $(N_q\text{-}A_q\text{-}1)(NV\text{-}A_q)$ degrees of freedom[6], where

$$F = \sum_{k=1}^{NV} \frac{\epsilon_{pk}^{(q)2}}{(NV - A_q)} \Bigg/ \sum_{k=1}^{NV} \sum_{p=1}^{N_q} \frac{\epsilon_{pk}^{(q)2}}{(N_q - A_q - 1)(NV - A_q)} \qquad (5)$$

The denominator in the above expression represents the total residual variance, S_0^2, and is a measure of the ability of the class model to accurately represent the data. Errors in classification were realized when (1) a test object was not assigned to its appropriate class or was assigned to no class at all or (2) when objects used in modeling one class were incorrectly assigned to another class. In this manner, each object was fitted to every class model marking the total number of tests 640 under Scheme I (80 objects fitted to 8 classes) and 480 under Scheme II (80 objects fitted to 6 classes).

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 interclass distances and variable discriminatory power were written in BASIC programming language.

RESULTS AND DISCUSSION

Assignment of resonances in peracetylated carbohydrate derivatives.—The strategy used in assigning resonances to the protons of the carbohydrate backbone, the acetoxyl methyl protons, and the carbonyl carbons can be viewed as a two-step process. Initially, a variety of homonuclear correlation spectroscopies are used to assign the protons of the carbohydrate backbone. In past studies this has been exclusively done using the COSY sequence[1–4]. In the present study a variety of other experiments have been used including COSY optimized for the observation of long-range couplings (DCOSY) and HOHAHA[27,28]. The HOHAHA experiment provides through-space correlations across several bonds in the same residue, providing a check on assignments made via three bond couplings using the COSY experiment[29,30]. An example of the utility of the experiment is shown in Fig. 1, the HOHAHA spectrum of the peracetylated tetrasaccharide β-D-Gal-(1 → 4)-β-D-GlcNAc-(1 → 3)-β-D-Gal-(1 → 4)-β-D-Glc-(1 → O)-Me (**13**). In some cases resonances for all the pyranosyl ring protons appear as subspectra of the two-dimensional contour map. This is illustrated in Fig. 1 where four off-diagonal contours arising from the H-2–H-5 resonances appear on the horizontal lines drawn through the diagonal contour representing the anomeric proton resonances of the peracetylated glucosyl and 2-acetamido-2-deoxyglucosyl residues (residues A and

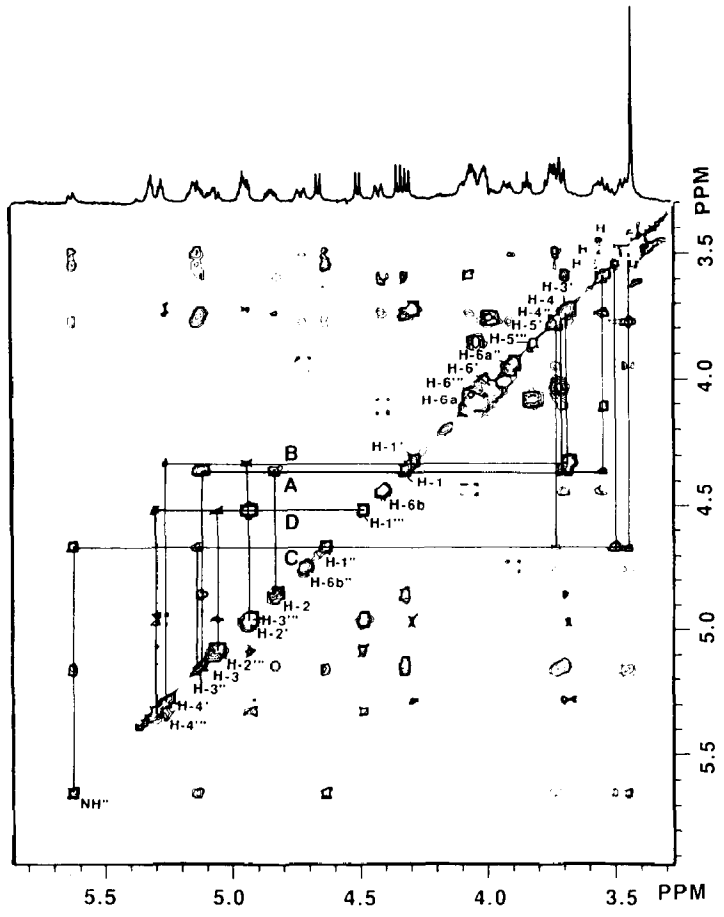

Fig. 1. The two-dimensional HOHAHA spectrum of β-Gal-(1 → 4)-β-GlcNAc-(1 → 3)-β-Gal-(1 → 4)-β-Glc-(1 → O)-Me, peracetylated with [1,1′-^{13}C]acetic anhydride (**13**). Horizontal lines through the anomeric proton resonances lying along the diagonal are drawn to denote subspectra for each of the four residues, where residue A is the derivatized methyl glucopyranoside. The normal ^{1}H NMR spectrum is shown at the top of the figure.

C). By comparison with the COSY spectrum, showing only connectivities between neighboring protons, assignments can be made directly. Once the H-5 resonance is assigned, the H-6 and H-6′ assignments may be made from either the COSY spectrum or by drawing a horizontal line through the H-5 contour in the HOHAHA spectrum. In the case of a galactose residue, such as the two in **13**, no connectivity is observed between H-4 and H-5 either in the HOHAHA spectrum or the COSY spectrum. The *gauche* configuration of these protons minimizes their through bond coupling and their mutual cross-relaxation rate in the rotating frame[31]. The H-5 resonances for these residues were finally assigned from the DCOSY spectrum of **13**, where connectivity was apparent between H-5 and H-4.

Once the proton spectrum of a peracetylated carbohydrate has been completely assigned, the ^{13}C-substituted carbonyl carbons can be assigned using heteronuclear

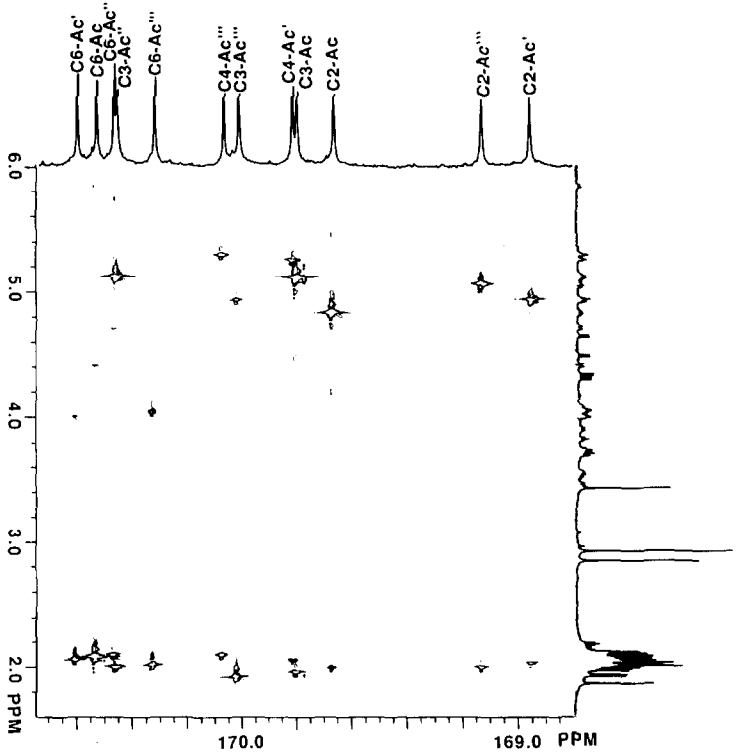

Fig. 2. The two-dimensional COLOC spectrum of **13**. Normal ^{13}C and ^{1}H NMR spectra are shown along the horizontal and vertical axes. Notations for carbonyl carbons are similar to those used in Table 1.

correlation spectroscopy. Because the residual long-range proton coupling present in the more sensitive proton-detected HMBC experiment[32] can often complicate proton correlations to overlapping carbon resonances, we have chosen instead to use conventionally detected carbon–proton correlation experiments. Fig. 2 demonstrates how the COLOC experiment[28] can be used to correlate previously assigned pyranosyl ring proton resonances and heretofore unassigned acetoxymethyl proton resonances to nearest neighbor carbonyl resonances. The resonance assignments determined from correlation experiments are summarized in Table I.

Selection of variables and K-nearest neighbor classifications.—The overall objective of this and previous studies has been to use the NMR data to correctly classify residues contained in a previously unseen peracetylated carbohydrate derivative[3–5]. Detailed classifications might use data such as those listed in Table I as a basis for determining residue type, anomeric ring form, and the position of glycosidic substitutions to nearest neighbor residues. As a first step in carrying out such classifications, each residue within a parent structure is characterized by a vector of NMR parameters taken from one- and two-dimensional NMR experiments. The variables composing the vector include the coupling constant, $J_{H\text{-}1,H\text{-}2}$, and the

TABLE I

Summary of NMR chemical shift and coupling constant data on peracetylated carbohydrate derivatives (**1–13**) *

Acetoxyl	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	$J_{H\text{-}1,H\text{-}2}$
	β-Gal-(1 → 3)-β-Gal-(1 → O)-Me (**1**)				α-Gal-(1 → 3)-α-Gal-(1 → O)-Me (**2**)			
C-1		4.27		7.64		4.29		3.13
C-2	168.99	5.14	2.06		170.29	5.02	2.09	
C-3		3.80				4.07		
C-4	170.00	5.35	2.08		169.79	5.38	2.09	
C-5		3.78				4.25		
C-6	170.56	4.07, 4.11	2.03		170.38	4.07	2.00	
C-1′		4.54		7.38		5.22		3.43
C-2′	169.17	5.04	1.97		170.03	5.27	2.01	
C-3′	170.17	4.90	1.92		169.95	5.20	1.91	
C-4′	170.33	5.31	2.12		170.20	5.44	2.09	
C-5′		3.82				4.42		
C-6′	170.40	4.07, 4.14	2.01		170.37	4.00, 4.27	2.03	
	β-Gal-(1 → 4)-α-Man-(1 → O)-Ac (**3**)				β-Gal-(1 → 4)-β-Man-(1 → O)-Ac (**4**)			
C-1	168.18	5.98	2.12	2.29	168.20	5.77	2.05	3.37
C-2	169.42	5.19	2.17		169.90	5.41	2.16	
C-3	169.33	5.31	2.02		169.05	5.13	2.02	
C-4		3.92				3.87		
C-5		3.92				3.73		
C-6	170.34	4.12, 4.36	2.08		170.38	4.15, 4.36	2.08	
C-1′		4.50		7.93		4.51		7.82
C-2′	169.18	5.11	2.03		169.20	5.08	2.01	
C-3′	170.00	4.93	1.94		170.00	4.95	1.93	
C-4′	170.04	5.30	2.12		170.03	5.30	2.12	
C-5′		3.86				3.84		
C-6′	170.29	4.00, 4.14	2.02		170.28	4.00, 4.14	2.02	
	β-Gal-(1 → 4)-β-Glc-(1 → O)-Me (**5**)				β-Gal-(1 → 3)-α-GalNAc-(1 → O)-Ac (**6**)			
C-1		4.39		7.69	168.67	6.27	2.11	1.39
C-2	169.62	4.88	2.05			4.60		
C-3	169.70	5.20	2.05			4.01		
C-4		3.81			169.92	5.42	2.10	
C-5		3.61				4.19		
C-6	170.33	4.05, 4.06	2.02		170.58	3.99, 4.14	2.02	
C-1′		4.49		7.84		4.70		7.98
C-2′	168.99	5.09	2.04		169.91	5.19	2.05	
C-3′	169.97	4.98	2.01		170.11	4.99	1.94	
C-4′	170.06	5.34	2.10		170.15	5.38	2.14	
C-5′		3.88				3.94		
C-6′	170.26	4.08, 4.10	2.06		170.47	4.17, 4.17	2.03	
	β-Gal-(1 → 6)-α-GlcNAc-(1 → O)-Ac (**7**)				β-Gal-(1 → 6)-β-GlcNAc-(1 → O)-Ac (**8**)			
C-1	168.61	6.12	2.12	3.25	169.33	5.62	2.04	8.78
C-2		4.40				4.20		
C-3	171.66	5.16	1.98		171.05	5.06	1.97	
C-4	169.28	4.94	1.98		169.33	4.93	1.99	
C-5		3.87				3.65		
C-6		3.38, 3.85				3.49, 3.85		
C-1′		4.45		7.95		4.49		7.98
C-2′	169.62	5.14	2.00		169.57	5.18	2.00	

(continued)

TABLE I (continued)

Acetoxyl	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-3′	170.04	4.98	1.91		170.08	4.99	1.91	
C-4′	170.15	5.36	2.07		170.17	5.35	2.07	
C-5′		3.90				3.53		
C-6′	170.34	4.15, 4.10	1.99		170.36	4.03, 4.11	1.99	
	β-GalNAc-(1 → 3)-α-Gal-(1 → O)-Me (**9**)				β-GlcNAc-(1 → 3)-β-Gal-(1 → O)-Me (**10**)			
C-1		4.92		3.40		4.29		7.73
C-2	170.28	5.15	2.16		169.50	5.09	2.12	
C-3		4.20				3.82		
C-4	170.03	5.43	2.16		169.83	5.35	2.12	
C-5		4.15				3.79		
C-6	170.54	4.02, 4.15	2.08		170.76	4.10	2.12	
C-1′		4.93		7.73		5.62		8.25
C-2′		3.67				3.29		
C-3′	170.32	5.39	1.99		170.42	5.50	1.99	
C-4′	170.30	5.33	1.99		169.50	5.01	1.93	
C-5′		3.88				3.60		
C-6′	170.43	4.07, 4.15	2.07		170.64	4.05, 4.27	2.10	
	α-GalNAc-(1 → O)-Ac (**11**)				β-GalNAc-(1 → O)-Ac (**12**)			
C-1	168.80	6.24	2.17	3.30	169.62	5.80	2.15	8.07
C-2		4.71				3.88		
C-3	171.15	5.25	2.02		170.81	5.18	2.04	
C-4	170.21	5.44	2.17		170.19	5.39	2.18	
C-5		4.25				4.11		
C-6	170.37	4.05, 4.21	2.03		170.43	4.05, 4.21	2.05	
	β-Gal-(1 → 4)-β-GlcNAc-(1 → 3)-β-Gal-(1 → 4)-β-Glc-(1 → O)-Me (**13**)							
C-1		4.35		7.90				
C-2	169.67	4.83	2.00					
C-3	169.41	5.12	1.97					
C-4		3.71						
C-5		3.56						
C-6	170.54	4.09, 4.41	2.08					
C-1″		4.64		7.90				
C-2″		3.54						
C-3″	170.47	5.14	2.01					
C-4′		3.75						
C-5″		3.45						
C-6″	170.48	3.93, 4.71	2.10					
C-1′		4.31		8.02				
C-2′	168.96	4.95	2.03					
C-3′		3.69						
C-4′	169.82	5.26	2.06					
C-5′		3.73						
C-6′	170.61	4.00, 4.01	2.06					
C-1‴		4.49		7.89				
C-2‴	169.14	5.05	2.00					
C-3‴	170.02	4.92	1.92					
C-4‴	170.08	5.30	2.10					
C-5‴		3.84						
C-6‴	170.33	4.05, 4.06	2.02					

* All chemical shifts are with respect to Me_3Si used as an internal standard. The abbreviations used are ^{1}H-PR for the pyranosyl ring proton, ^{1}H-AcMe for the acetoxyl methyl proton, and ^{13}C-Ac for the acetoxyl carbonyl carbon.

resonance chemical shifts of the pyranosyl ring protons, the H-6 methoxy protons, the acetoxymethyl protons and the acetoxy carbonyl carbons, for a total of fifteen variables (eq *1*). A relatively simple method of classification is the KNN method[3–5,6,13,14], where unknown residues are classified according to their Euclidean distance to their nearest neighbor or two nearest neighbors in a data set obtained from compounds of known structure. Whether or not a classification is made correctly will depend on the defined classes of the data set and the similarity of variables between the test residue and those of residues already present in the data set. In the present investigation each member residue of the data set is treated as a test residue and is classified by the KNN method according to one of two selected classification schemes. Using the first of these schemes (Scheme I), it was determined if each test residue could be correctly classified into one of eight classes made up from the remaining data set members. Seven of these eight classes consisted of residues selected on the basis of their structure and anomeric ring configuration (α-D-glucoses, β-D-glucoses, α-D-mannoses, α-D-galactoses, β-D-galactoses, 2-acetamido-2-deoxy-α-D-glucoses and 2-acetamido-2-deoxy-β-D-glucoses). The eighth class was selected only on the basis of structure (2-acetamido-2-deoxy-α,β-D-galactose) due to the limited number of members of each anomeric form within the class (two each). According to the second classification scheme (Scheme II), the two classes representing 2-acetamido-2-deoxy-α- and β-D-glucose were combined into a single class as were the two α- and β-D-galactose classes, resulting in a total of six classes.

In previous studies we have found that the overall effectiveness of the KNN classification method can be improved if variables having little value in discriminating between classes are eliminated[3,4]. This arises because these variables provide little information and add "noise" to the classification method. A quantitative description of the relative importance a variable in discriminating between pairs of classes can be realized using methods contained in the SIMCA pattern recognition method[3–7,9–12]. Accordingly, each class was modeled by a single principal component pointing in the direction of greatest variance of the data (eq *2*). A measure of the discriminatory power of a variable was evaluated from the residuals obtained from the fit of the objects of one class to the principal component of another class (eq *3*). The average discriminatory power of each variable, obtained for the average of such pairwise interactions, was used in evaluating their relative importance. When the entire data set was divided into eight classes, their relative importance was found to be δ(C-2 Ac) > δ(H-4) > δ(H-5) > $J_{H\text{-}1,H\text{-}2}$ > δ(H-2) > δ(C-2 AcMe) > δ(H-1) > δ(C-4 Ac) > δ(C-6 AcMe) > δ(C-6 Ac) > δ(C-4 AcMe) > δ(H-6) > δ(H-3) > δ(C-3 Ac) > (C-3 AcMe). Data sets were then formed which were described by the ten, eight, five and four most important variables (DP-10, DP-8, DP-5, and DP-4). For comparison, two additional data sets were formed having as their basis some or all of the pyranosyl ring proton chemicals shifts and $J_{H\text{-}1,H\text{-}2}$ (SP-3 and SP-6). These data sets were formed using variable discriminatory power as a guiding rather than an absolute criterion. Results for KNN classifica-

TABLE II

Results of K-nearest-neighbor calculations

Data Set	Variables	Number misassigned (% correctly classified)			
		Classification Scheme I		Classification Scheme II	
		1-KNN	3-KNN	1-KNN	3-KNN
VAR-15	Complete data set	12(85)	19(76)	10(87)	15(81)
DP-10	H-1, H-2, H-4, H-5, C-2 Ac, C-2 AcMe, C-4 Ac, C-6 Ac, C-6 AcMe, $J_{H-1,H-2}$	13(84)	16(80)	9(89)	9(89)
DP-8	H-1, H-2, H-4, H-5, C-2 Ac C-2 AcMe, C-4 Ac, $J_{H-1,H-2}$	8(90)	11(86)	3(96)	7(92)
DP-5	H-2, H-4, H-5, C-2 Ac, $J_{H-1,H-2}$	13(84)	17(79)	10(87)	15(81)
DP-4	H-4, H-5, C-2 Ac, $J_{H-1,H-2}$	12(85)	14(82)	9(89)	13(84)
SP-6	H-1, H-2, H-3, H-4, H-5, $J_{H-1,H-2}$	14(82)	24(70)	13(84)	20(75)
SP-3	H-1, H-2, $J_{H-1,H-2}$	11(86)	16(80)	10(87)	13(84)

tions using classification schemes discussed above are shown in Table II. As has been the case in previous studies, there are fewer incorrect classifications of test residues when the nearest neighbor rather than the nearest two or three neighbors is used as a classification criterion. This arises primarily due to the limited number of residues in each class. It is also clear from the data that fewer errors are made when residues are classified with respect to six rather than eight classes. The difference between these results has as its basis the number of times an α-D-galactose or an 2-acetamido-2-deoxy-α-D-glucose is mistakenly classified into the class of its respective β anomer. Again, this most likely arises due to the scarcity of similar members in both of the α anomeric classes. As was seen previously, the elimination of variables having respectively low discriminatory power results in fewer misclassifications[3,4]. This trend appears to be independent of how classes are selected. When the basis set of variables is reduced to less than eight variables, variables important in discriminating between classes are lost, and the number of misclassifications of test residues increases.

Some insight into the reasons for the greater number of misclassifications for data sets having as their basis fewer than eight variables can be gained from the principal component plot shown in Fig. 3. These graphs show the entire data set along the two directions of greatest variance of the data[6–9]. When the eight variables most important for interclass separation are used as a basis set (Fig. 3A), residues having similar structures, for the most part, cluster in different regions of the plot. Careful study of the plot shows that often subclusters of two or more residues appear within a cluster of data representing a single residue type. These subgroups arise from residues with overall structures similar to the rest of the group but having differing detailed structures. This plot may be contrasted with the principal component plot of the data set when only the chemical shifts of H-1 and H-2 and the coupling constant $J_{H-1,H-2}$ are used as a variable basis set (Fig. 3B). Because these variables are most sensitive to the anomeric ring form, there is a

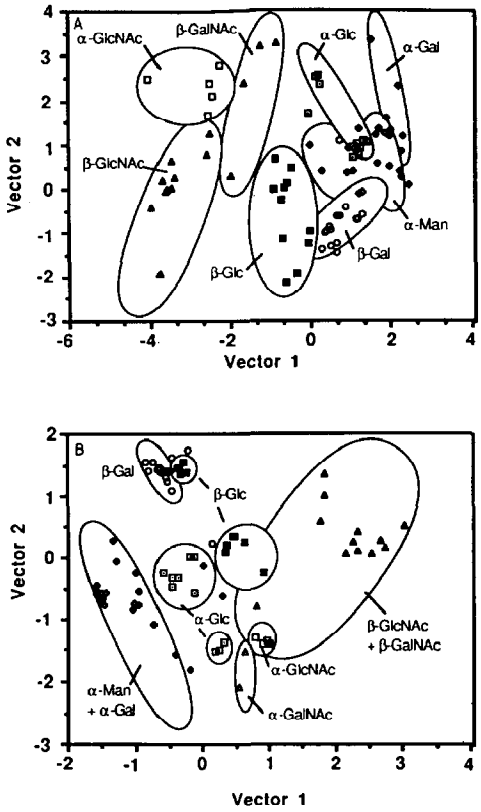

Fig. 3. Principal component plots of the entire 8-residue data set. The two axes, formed from admixtures of NMR variables included in the basis set, point in the directions of greatest variance in the data. (A) Principal components were constructed from eight of the 15 NMR variables having the greatest discriminatory power (DP-8) (B) Principal components were constructed from $J_{H\text{-}1,H\text{-}2}$ and the H-1 and H-2 resonance shifts (SP-3). Symbols used in the plot to denote the α-D-glucose (□), β-D-glucose (■), α-D-galactose (♦), β-D-galactose (○), α-D-mannose (◇), 2-acetamido-D-glucose (□), 2-acetamido-β-D-glucose (▲) and 2-acetamido-α-(or β-)-D-galactose (△) residues.

major division in the plot with residues having the β ring form clustering in the upper right and those having the α ring form appearing in the lower left. More importantly, the absence of chemical shift information near to sites of structural differences between classes results in having the two classes appear the same. Hence, 2-acetamido-2-deoxy-β-D-glucose residues appear in the same region of the plot as 2-acetamido-2-deoxy-β-D-galactose residues due to the absence of chemical shift information about the C-4 pyranosyl ring site. Similarly, both anomeric forms of D-glucose cluster in the same region of the plot as those of D-galactose. The absence of variables in the data set which represent nuclei near to sites of structural variation may account for many of the misclassifications summarized in Table II.

Classification of residues using SIMCA class modeling.—Previous results based on data sets having fewer member residues have indicated that residues may be

TABLE III

Results of SIMCA calculations

Classes	Variable set				
	VAR-15	DP-10	DP-8	DP-5	SP-6
	Classification Scheme I				
	Number misassigned (% correctly assigned)				
α-Glc	0	1	4	4	0
α-Gal	0	14	12	1	20
β-Gal	0	0	0	2	2
α-Man	6	5	7	9	6
β-Glc	0	1	6	11	1
α-GlcNAc	0	0	0	0	0
β-GlcNAc	0	0	2	9	44
α,β-GalNAc	0	0	0	0	32
Totals	6 (99.1)	21 (96.8)	31 (95.2)	35 (94.4)	105 (83.6)
	Classification Scheme II				
	Number misassigned (% correctly assigned)				
α-Glc	0	1	4	4	0
α,β-Gal	0	1	8	9	11
α,β-Man	8	6	6	11	18
β-Glc	0	2	7	13	2
α,β-GlcNAc	2	3	4	4	56
α,β-GalNAc	0	0	0	0	33
Totals	10 (98.0)	13 (97.3)	29 (94.0)	41 (91.5)	120 (75.0)

more accurately classified if the basis for classification is statistical in nature rather than the more simplistic KNN approach[3]. Using the SIMCA method, each of the classes in the data set is independently modeled using principal components[6–12]. The method is similar to a series expansion of a function about a point, where any number of principal components up to the total number of variables may be used in the expansion. Typically modeling is carried out using some of the objects contained in a class training set, and the validity of the model can be tested on the remaining objects belonging to the same class (test objects). In the present case, twelve or fifteen of the 80 total residues served as test objects when classification was carried out according to Schemes I or II, respectively (15 or 18% of the data set). Each class was modeled using one principal component. Both the test objects and those used in modeling are classified using as a basis for classification an F test at the 90% confidence level (eq 5). Errors in classification were realized when (1) a test object was not assigned to it appropriate class or assigned to no class at all or (2) when objects used in modeling one class were incorrectly assigned to another class. In this manner, every object was fitted to every class model.

A summary of results of the SIMCA calculations is given in Table III, using as a basis for modeling some of the same reduced variable sets as were used for the KNN classification. These results show that, when the full compliment of fifteen variables are used for class modeling, greater than 98% of the residues are

classified correctly under either of the two classification schemes. Under these conditions misclassifications most frequently arose from instances in which α-D-mannose residues were misclassified as α-D-glucose residues. None of the residues misclassified were test residues but were instead those used in modeling their own class. The results also show that in general as the number of variables used in class modeling is reduced, the number residues misclassified tends to increase. This trend is in contrast to the KNN results which showed optimum classification was achieved with an eight variable basis set. However, even when the basis set is reduced to those five variables having the greatest discriminatory power, the number of residues classified correctly is greater than 94 or 91%, respectively, under the Scheme I and II classification methods. The number of misclassified residues markedly increases when the criterion used for selecting the reduced variable basis set is not based on variable discriminatory power. This is exemplified by the comparatively poor SIMCA results found using only proton chemical shifts and coupling constants as a basis for class modeling (SP-6). In these cases the majority of errors arose when 2-acetamido-2-deoxy-D-glucoses and 2-acetamido-2-deoxy-D-galactose residues were mistakenly classified as D-glucose or D-galactose residues. Similar misclassifications were not seen when other basis set were used due to the unique chemical shifts assigned to C-2 Ac carbonyl resonance and C-2 AcMe proton resonance. In the absence of these unique shifts parameters characterizing 2-acetamido-2-deoxy-D-glucose residues appear quite similar to those of their corresponding sugars, not having a 2-acetamido-2-deoxy substitution at C-2.

CONCLUSIONS

Our results show that the KNN method is more successful in classifying residues correctly when the number of residues in each class is maximized at the expense of some structural diversity within the classes. As was previously shown[3,4], the results of the classification can be improved if those variables not important in discriminating between classes are not included in the nearest-neighbor calculations. In contrast to these results, when SIMCA is used as a basis for making classifications, class homogeneity appears to take precedence over the number of residues within each class. Apparently homogeneous classes are more accurately modeled by SIMCA than are larger classes having greater structural diversity. Furthermore, any elimination of variables from the basis set used in modeling seems to have adverse effects on the ability of SIMCA to model these classes and ultimately results in a greater number of misclassifications.

In conclusion, our results indicate that the SIMCA method is superior to the KNN method for classifying peracetylated carbohydrate residues according to their structure, particularly when the data set is small and the number of structurally distinct classes is large. Both methods were most successful in classifying residue structures on the basis of NMR variables when some or all of the carbonyl carbon or acetoxymethyl proton shifts were included in the data set. This result supports the use of comprehensive NMR data to identify oligosaccharide substructures. In

comparison, relatively poor results were obtained when only pyranosyl proton shifts and coupling constants were used to identify oligosaccharide residue structures.

ACKNOWLEDGEMENTS

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.*, (1991) in press.
5 W.J. Goux, in J.W. Finley, S.J. Schmidt, and A.S. Serrianni (Eds.), *NMR Applications in Biopolymers (Basic Life Sciences Ser.,* Vol. 56), Plenum Press, New York, 1990, pp. 47–62.
6 M.A. Sharaf, D.L. Illman, and B.R. Kowalski, *Chemometrics (Chemical Analysis Ser.*, Vol. 82), John Wiley and Sons, New York, 1986, pp. 179–296.
7 C. Albano, G. Blomquist, W. Dunn 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.
8 B.R. Kowalski, *Anal. Chem.*, 47 (1975) 1152A–1162A.
9 S. Wold and M. Sjostrom, *ACS Symp. Ser.*, 52 (1976) 243–252.
10 M. Sjostrom and U. Edlund, *J. Magn. Res.*, 25 (1977) 285–297.
11 U. Edlund and S. Wold, *J. Magn. Res.*, 37 (1980) 183–194.
12 S. Wold, *Pattern Recog.*, 8 (1976) 127–139.
13 B.R. Kowalsky and C.F. Bender, *Anal. Chem.*, 44 (1972) 1405–1411.
14 P.C. Jurs, *Science*, 232 (1986) 1219–1224.
15 J. Montreuil, *Adv. Carbohydr. Chem. Biochem.*, 37 (1980) 157–223.
16 D.A. Cumming, R.N. Shah, J.J. Krepinsky, A.A. Grey, and J.P. Carver, *Biochemistry*, 26 (1987) 6655–6676.
17 J.F.G. Vliegenthart, H. van Halbeek, and L. Dorland, *Pure Appl. Chem.*, 53 (1981) 45–77.
18 J.F.G. Vliegenthart, L. Dorland, and H. van Halbeek, *Adv. Carbohydr. Chem. Biochem.*, 41 (1983) 209–374.
19 D.R. Anderson and W.J. Grimes, *Anal. Biochem.*, 146 (1985) 13–22.
20 E.F. Hounsell, D.J. Wright, A.S.R. Donald, and J. Feeney, *Biochem. J.*, 223 (1984) 129–143.
21 E.F. Hounsell and D.J. Wright, *Carbohydr. Res.*, 205 (1990) 19–29.
22 J. Dabrowski, U. Dabrowski, P. Hanfland, M. Kordowicz, and W.E. Hull, *Magn. Reson. Chem.*, 24 (1986) 59–69.
23 E. Berman, U. Dabrowski, and J. Dabrowski, *Carbohydr. Res.*, 176 (1988) 1–15.
24 M. Ikura and K. Hikichi, *Carbohydr. Res.*, 163 (1987) 1–8.
25 S.W. Homans, R.A. Dwek, J. Boyd, N. Soffe, and T.W. Rademacher, *Proc. Natl. Acad. Sci. U.S.A.*, 84 (1987) 1202–1205.
26 B. Meyer, T. Hansen, D. Nute, P. Albersheim, A. Darvill, W. York, and J. Sellers, *Science*, 251 (1991) 542–544.
27 A. Bax, *Two-Dimensional Nuclear Magnetic Resonance in Liquids*, Delft University Press, Delft, Netherlands, 1982, pp. 50–98.
28 G.E. Martin and A.S. Zektzer, *Two-Dimensional NMR Methods for Establishing Molecular Connectivity*, VCH, New York, 1988, pp., 58–347.
29 R.A. Byrd, W. Egan, M.F. Summers, and A. Bax, *Carbohydr. Res.*, 166 (1987) 47–58.
30 L. Lerner and A. Bax, *Carbohydr. Res.*, 166 (1987) 35–46.
31 J.H. Noggle and R.E. Schirmer, *The Nuclear Overhauser Effect*, Academic Press, New York, 1971, pp. 22–43.
32 A. Bax and M.F. Summers, *J. Am. Chem. Soc.*, 108 (1986) 2093–2094.