dimerization reaction (eq 6) is uncertain, but bridged structures containing one OH^- bridge and another containing one OH^- and one deprotonated amide nitrogen bridge have been proposed[6] for the analogous glycylglycine complex, $Cu_2(H_{-1}Glygly)_2(OH)^-$. Since K_D for the formation of these dimers is the same (Table I) for all isomers of Leu-Leu, it appears that there is relatively little peptide rearrangement during dimer formation and also there is probably little contact between the two peptide ligands in the dimer. Thus the structure with one OH^- group bridging the two $Cu(H_{-1}L)$ residues *via* the Cu(II) atoms appears to be the most probable.

Acknowledgment. This investigation was supported by National Institutes of Health Research Grant GM12626 from the National Institute of General Medical Sciences.

Heuristic Pattern Recognition Analysis of Carbon-13 Nuclear Magnetic Resonance Spectra

Charles L. Wilkins,* Robert C. Williams, Thomas R. Brunner, and Patrick J. McCombie

Contribution from the Department of Chemistry, University of Nebraska—Lincoln, Lincoln, Nebraska 68508. *Received February 23, 1974*

Abstract: The first application of linear discriminant function analysis to experimental proton noise-decoupled ^{13}C high-resolution nuclear magnetic resonance spectra is reported. Results of various preprocessing methods are discussed and the implications with respect to the usefulness of the present approach for structural elucidation problems are considered. It is shown that the analysis of nmr data *via* a learning machine approach is comparable in efficacy to previous studies where mass and infrared spectral data were interpreted in the same way for the purpose of answering structural questions.

Since the use of the heuristic pattern recognition technique called the "learning machine"[1] method was introduced to chemical data analysis by Isenhour and coworkers,[2] its feasibility as a general approach to the interpretation of masses of experimental data has been studied extensively.[3-6] Among the structural elucidation techniques which have been examined most in this way are mass spectrometry[7-9] and infrared spectrometry.[10,11] We report here the first application of linear discriminant function analysis to natural abundance noise-decoupled ^{13}C nuclear magnetic resonance data. Roberts has suggested that the enormous sensitivity of ^{13}C chemical shifts to structural changes should make this technique a far more useful tool for the investigation of structure than proton nmr.[12] Because of the availability of instrumentation for relatively routine determination of high-resolution natural abundance ^{13}C nmr spectra, it seems imperative that rapid effective means of interpreting such data be developed.

In this paper, an entirely new approach to interpretation of ^{13}C nmr spectra, proceeding directly from spectrum to structural information and circumventing the detailed assignment of chemical shifts and coupling constants, is outlined.

Experimental Section

Data Base. As a data base for the study we have used a recently published collection of ^{13}C nmr spectra containing a total of 500 spectra measured on two different instruments and in eight different spectral solvents.[13] Chemical shifts were referenced to tetramethylsilane and, for the most part, covered a range of 200 ppm. Eighty of the spectra were obtained in the continuous-wave mode, the remainder were determined using Fourier transform operation. Intensities were digitized manually and added to the original structure-coded, peak frequency list contained in Johnson and Jankowski's collection.[13]

Computation Method. Binary pattern classification using a simple error correction feedback method[5] and various preprocessing methods was employed to analyze the coded spectral data. Programs were written in Fortran IV, using algorithms described below, and all computations were carried out using an IBM 360/65 computer. A typical computation including preprocessing, feature selection and development of a final weighting vector required between 1 and 3 min of central processor time.

Results and Discussion

Briefly, the analytical approach is to represent the ^{13}C nmr spectra as points in pattern space and then to find hyperplanes (linear discriminant functions) which separate them into binary subsets. Such decision surfaces may be developed for any desired binary choice (*e.g.*,

[1] N. J. Nilsson, "Learning Machines," McGraw-Hill, New York, N. Y., 1965.
[2] P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, **41**, 21 (1969).
[3] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **94**, 5632 (1972).
[4] B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, **96**, 916 (1974).
[5] T. L. Isenhour and P. C. Jurs, *Anal. Chem.*, **43**, No. 10, 20A (1971).
[6] L. B. Sybrandt and S. P. Perone, *Anal. Chem.*, **44**, 2331 (1972).
[7] J. B. Justice and T. L. Isenhour, *Anal. Chem.*, **46**, 223 (1974).
[8] P. C. Jurs, *Anal. Chem.*, **43**, 22 (1971).
[9] B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilly, *Anal. Chem.*, **41**, 1949 (1969).
[10] B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilly, *Anal. Chem.*, **41**, 1945 (1969).
[11] R. W. Liddell, III and P. C. Jurs, *Appl. Spectrosc.*, **27**, 371 (1973).
[12] H. J. Reich, M. Jautelat, M. T. Messe, F. J. Weigert, and J. D. Roberts, *J. Amer. Chem. Soc.*, **91**, 7445 (1969).
[13] L. F. Johnson and W. C. Jankowski, "Carbon-13 NMR Spectra," Wiley, New York, N. Y., 1972. The computer-readable spectral data were used with permission of the authors and the publisher.

separation of spectra into two classes, those of ketones and those of nonketones). In order to categorize a particular spectrum (pattern) as belonging to one of two possible classes, the decision function is applied and assignment made based on whether the scalar (S) thus obtained has a value less than zero or not. In the present research the spectra were each represented by a series of 200 data values (corresponding to sums of intensities or transformed intensities) and the dimension was augmented by an additional value, arbitrarily selected as 1, to ensure the decision hyperplane would pass through the origin of the resulting 201-coordinate hyperspace.[14] Thus, each spectrum could be viewed as a 201-element vector or, equivalently, as a single point in 201-dimensional hyperspace. A linear error correction feedback method was employed, in combination with a simple selection algorithm, to develop decision planes for a number of structural categorizations.[11] Equations 1–3 summarize the computational approach,

$$S = \sum_{i=1}^{201} W_i \cdot \mathbf{X}_i \tag{1}$$

$$W_i' = W_i + c\mathbf{X}_i \tag{2}$$

$$c = -S \Big/ \sum_{i=1}^{201} \mathbf{X}_i \cdot \mathbf{X}_i \tag{3}$$

where the W_i are the coefficients of the decision hyperplane and the $\mathbf{X}_i$ the 201-element vector corresponding to a spectrum. The subscripts 1–200 correspond to the 1-ppm spectral resolution elements used to represent the original data. Within each of these intervals, all peak intensities were summed to yield the raw data values for studies utilizing intensities or, for binary coded data (*vide infra*), were set equal to the number of peaks appearing in the particular interval. The first resolution element (1) contained, in addition, all data from peaks appearing at higher field than TMS. The last resolution element (200) contained information on all spectral peaks appearing with shifts greater than 200 ppm. Coding in this manner amounts to approximately a tenfold degradation of the quoted spectral resolution in the original data. For error correction feedback, modified decision hyperplanes (W') are developed using the correction algorithm stipulated in eq 2 and 3. This correction feedback is applied for each misclassified spectrum in a training set comprised of 400 of the original 500 spectra in an iterative fashion until either perfect convergence classification occurs or a predetermined number of feedbacks have taken place.[5]

Feature Elimination. For feature elimination (and, therefore, dimension reduction) a simple procedure is used. First, an arbitrary starting weight vector (W) initialized to all ones is applied to each of the 400 members of the training set, whose correct categories are known. This weighting vector is then improved as described above. When no further improvement in classification occurs, the computation is repeated using a starting **W** initialized to all minus ones. Those features (resolution elements) whose final weight vector coefficients change sign as a result of the change in initialization conditions are eliminated. The entire process is then repeated iteratively until no further feature elimination occurs. As a test of the efficacy of the weight vectors developed at each stage of the process, their ability to classify 100 members of the original data set (the "unknown" set) which were not included in the training set is determined.

Carbon-13 Nmr Preprocessing. *A priori*, there are no certain guiding principles to dictate a choice of preprocessing methods for an analysis such as that described here. That is not to say, however, that some logical choices cannot be made. Consideration of the physicochemical basis of the noise-decoupled ^{13}C nmr data reveals two essential facts which can serve as guides in selection of particular preprocessing approaches. First, due to large ratio of ^{13}C nmr shift range to line width for typical organic molecules, their noise-decoupled ^{13}C spectra tend to contain a single resonance for each carbon nucleus. This suggests a simple peak–no peak coding scheme might be worthwhile. Second, the peak intensity differences are sensitive monitors of the difference in carbon environment (due to influences of Overhauser effects and differing T_1's). Thus, an intensity-based coding scheme might also prove useful. Which is best may depend on the type of structural information sought and/or the relative importance of the two types of information, peak position and peak intensity, in making one class separable from another. These factors were empirically assessed in the present study. An intermediate approach, wherein intensities were coded according to which of five intensity intervals they fitted, was also examined. When intensity information coded to the nearest 1% was used, two types of normalization were explored. The first was to assign the *largest* peak in each spectrum the value 100 and to code the intensities of the remaining peaks relative to that base value (*i.e.*, to assign them values between 1 and 100). A second normalization approach was to first code peaks as described above and then to scale the resulting values so that each spectrum summed to 100 (*i.e.*, all spectra are weighted equally). Table I contains the results of all of these approaches for a variety of functional group categorization questions.

Several facts are readily apparent from Table I. First, heuristic pattern recognition is successful for this particular data base. The technique is most successful for those cases where the chemical class is rather precisely defined. Thus, better results are achieved for the specific functions, "aldehydes and ketones," than for the less specific "carbonyl," which includes these functions as subsets, but also includes acids, esters, amides, etc. Clearly, a better weight vector will be found if the property is more specific.

Second, there is no obviously superior preprocessing technique. Based on the average per cent correct prediction for a 200 feature data set, the binary and scaled normalized intensity techniques give better results for two functional group classes each, scaled absolute intensities for one class, and for two classes equally good results are found for more than one preprocessing technique. When unnecessary features are eliminated, the binary and scaled absolute intensity methods give the best results, but one technique may yield results 4–5% better than another in any given case. When the speed of convergence is also considered, the binary technique begins to look slightly superior to the other, since it converges most rapidly in five of the 200-feature sets, and four of the reduced feature cases. However,

[14] L. E. Wangen, N. M. Frew, and T. L. Isenhour, *Anal. Chem.*, 43, 845 (1971).

Wilkins, et al. / Interpretation of ^{13}C Nmr Spectra

Table I. Error Correction Feedback Training for Functional Group Identification from ^{13}C Nmr Spectra

			No feature elimination[b,c]		Feature elimination[c]		
Functional group	Method	No. in[a] total set	No. of feedbacks +/−	% correct unknown set +/−	Final no. retained	No. of feedbacks +/−	% correct unknown set +/−
Aldehyde and ketone	PNP[d]	29	74/41	99/99	17	23/19	100/100
	AI[e]		173/112	92/93	55	65/23	98/98
	SAI[f]		101/89	93/96	42	35/21	100/100
	NAI[g]		193/130	85/95	90	144/123	96/96
	SNAI[h]		69/51	100/100	17	31/17	100/99
Aliphatic alcohol	PNP	78	393/309	88/89	100	430/336	88/91
	AI		1725/1417	74/75	130	*i*	69/72
	SAI		566/507	79/83	112	*i*	85/85
	NAI		*i*/1467	74/79	126	*i*	79/75
	SNAI		657/660	86/83	91	637/626	86/85
Carbonyl (any C═O)	PNP	167	265/458	80/77	81	198/274	82/79
	AI		1217/1647	72/65	133	1001/1296	70/64
	SAI		791/1034	75/70	105	535/792	73/74
	NAI		1018/1663	73/71	148	*i*	69/70
	SNAI		423/621	74/74	98	275/390	75/77
Carboxylic acid	PNP	31	162/127	95/94	88	*i*	95/94
	AI		215/169	91/95	120	*i*	92/89
	SAI		200/191	95/93	63	*i*	96/95
	NAI		191/168	90/93	92	125/125	95/94
	SNAI		158/165	95/94	94	*i*	47/97
Alkyl bromide	PNP	18	346/101	92/97	96	*i*	89/96
	AI		395/197	91/96	112	*i*	94/94
	SAI		323/136	94/96	74	301/193	93/92
	NAI		311/137	94/95	105	217/73	96/94
	SNAI		270/113	92/94	99	*i*	95/96
Alkyl chloride	PNP	14	196/101	96/100	92	*i*	94/96
	AI		223/115	96/100	112	*i*	94/94
	SAI		178/102	96/100	78	112/86	99/100
	NAI		195/115	94/100	114	*i*	96/97
	SNAI		134/98	100/100	71	*i*	95/86
Phenyl	PNP	130	465/393	76/76	86	452/434	80/75
	AI		946/616	76/78	102	1678/669	81/81
	SAI		798/623	75/77	93	1322/530	81/81
	NAI		1165/847	78/78	109	1386/1040	80/79
	SNAI		613/585	78/78	87	539/*i*	81/78

[a] Total set of 500 spectra (see Experimental Section). [b] Using 400 spectra as the training set and 200 features. The remaining 100 spectra comprise the unknown set. [c] + refers to all ones initial weight vector; − refers to all minus ones initial weight vector. [d] Binary; 1 for peak, 0 for no peak. [e] Absolute intensity. [f] Scaled intensity, maximum = 5; total intensity varies. [g] Normalized intensity; total = 100/spectrum. [h] Scaled intensity, maximum = 5; total intensity normalized to 100/spectrum. [i] Maximum number of feedbacks was reached or not linearly separable.

it should be noted that this aspect is only of importance when weight vectors are being developed. Thereafter, it has no effect on the use of the vectors.

Third, feature elimination yields results equally as good as, and usually better than, those achieved using the full 200 feature data set. The savings in computation time are also substantial, provided some method of detecting mutually exclusive divergent spectra is included (*i.e.*, when the same several spectra are fed back repetitively for several iterations). The number of features eliminated also depends on the method of preprocessing. The three methods using the least peak height resolution (PN, SAI, SNAI) yield the fewest features.

Fourth, more rapid convergence is obtained when the weight vectors are initialized to all −1's. This result should be expected, as long as the particular feature sought comprises less than 50% of the data base, since a completely negative weight vector will cause the binary pattern classifier to yield an entire set of "no" results. Similarly, a weight vector initialized to all +1's should converge more rapidly when the sought-for feature appears in a majority of the spectra. It is worth noting that one way of judging the ability of the method to classify unknown compounds is, as Schecter and Jurs suggest,[15] to compare its predictive ability with the results obtained by always guessing the more populous category. In all but carboxylic acid and alkyl bromide cases, our results significantly exceed this figure. Certainly a larger, more representative data base would be expected to improve the situation for the two categories where this was not true.

The present study is compared with pattern recognition studies on other spectroscopic data in Table II.

Table II. Comparison of ^{13}C Nmr, Ir, and Mass Spectral ECF Training and Feature Selection

	Prediction %/features retained		
Functional group	Nmr	Ir[a]	MS[b]
Aldehydes and ketones	100/17	90/79[c]	
Aliphatic alcohols	90/100	96/86	89/65
Carbonyl	81/81	99/70	73/65
Phenyl	81/93	90/93	95/65

[a] Reference 11. [b] P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **42**, 1387 (1970). [c] Ketones only.

The results obtained from ^{13}C nmr data are comparable, both in features retained, and prediction per cent, with the other two methods.

[15] J. Schecter and P. C. Jurs, *Appl. Spectrosc.*, 27, 30 (1973).

Conclusions

We conclude that the present results clearly establish the possibility of using heuristic pattern recognition-based interpretation of ^{13}C nmr data as a structural elucidation tool. Our model study, using a set of spectra obtained under a rather broad range of experimental conditions, certainly suggests that the use of ^{13}C nmr in this way is comparable in speed, reliability, and specificity with the earlier infrared and mass spectral methods using the same approach. Encouraged as we are by the results thus far obtained, studies of the direct interpretation of untransformed digitized impulse response (free induction decay) data, as suggested by Kowalski and Reilly,[16] are under way. Furthermore, we are also examining the use of Hadamard transform[17] preprocessing of nmr data as an alternate method of preliminary data reduction. Spectral simulation *via* a related approach *bypassing derivation of chemical shift and coupling constant parameters* is also being examined. It is our belief that the promise of the present and related studies is that an integrated ^{13}C pattern recognition–Fourier nmr laboratory computer system is a realistic possibility. Thus, we are proceeding with plans for implementation of such a system which, ultimately, will provide the possibility of placing the experiment itself in the data interpretation feedback loop.

Acknowledgment. Support of this work under Grant GP-41515X from the National Science Foundation is gratefully acknowledged. We also thank Mr. William Jankowski and Dr. LeRoy Johnson for kindly supplying a computer listing of the spectral data file used.

[16] B. R. Kowalski and C. A. Reilly, *J. Phys. Chem.*, **75**, 1402 (1971).
[17] B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **45**, 2234 (1973).

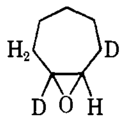
Conformational Analysis of Cycloheptene Oxide by ^{13}C and ^{1}H Nuclear Magnetic Resonance Spectroscopy

Kenneth L. Servis,*[1,2] Eric A. Noe,[2] N. Roy Easton, Jr.,[3] and F. A. L. Anet*[3]

Contribution from the Department of Chemistry, University of Southern California, Los Angeles, California 90007, and Contribution No. 3231 from the Department of Chemistry, University of California—Los Angeles, Los Angeles, California 90024. *Received December 26, 1973*

Abstract: Pmr spectra of cycloheptene-*1,3,3-d*$_3$ oxide in vinyl chloride solution were studied as a function of temperature from −150° to room temperature. Two different conformations in the ratio of 71:29 were observed at low temperatures. The free energy barrier ($\Delta G^{\ddagger}$) for conversion of the major conformation to the minor conformation is calculated to be 7.9 kcal/mol from a line-shape analysis at intermediate temperatures. Cmr spectra of cycloheptene oxide in a solution of $CHCl_2F/CHClF_2$ were obtained at temperatures in the range of −170 to −60°. In this case, two forms in the ratio of 60:40 were detected at low temperatures and a $\Delta G^{\ddagger}$ of 7.5 kcal/mol was obtained. It is suggested that cycloheptene oxide exists in two chair conformations of slightly different energies. Interconversion paths between these conformations are discussed.

Because the substituent sites in many seven-membered rings rapidly equilibrate by pseudorotation even at very low temperatures, relatively few low-temperature nmr studies of the conformations and barriers to conformational changes have been reported for rings of this size.[4] In the work described here, we have used nmr spectroscopy to study the conformational equilibration of cycloheptene oxide and its deuterated derivative, cycloheptene-*1,3,3-d*$_3$ oxide (**1**). As in cycloheptene, the pseudorotation pathway in the chair form of this epoxide is precluded by the restriction of the C_7–C_1–C_2–C_3 dihedral angle to a value near 0°, thus removing one possible conformational change.



1

Results

The deuterated epoxide, **1**, was synthesized from cycloheptanone by the route outlined in Scheme I. The compound was purified by preparative vpc[5] and identified by its pmr spectrum and by comparison of its vpc retention time with that of unlabeled cycloheptene oxide prepared from cycloheptene and *m*-chloroperbenzoic acid. This unlabeled cycloheptene oxide was also used for the cmr experiments.

The pmr spectrum at +30° for the proton at C-2 of **1**

[1] Alfred P. Sloan Research Fellow, 1969–1971.
[2] University of Southern California.
[3] University of California—Los Angeles.
[4] (a) E. S. Glazer, Ph.D. Thesis, California Institute of Technology, Pasadena, Calif., 1966; (b) J. D. Roberts, *Chem. Brit.*, 529 (1966); (c) R. Knorr, C. Ganter, and J. D. Roberts, *Angew. Chem.*, **79**, 577 (1967); (d) M. St. Jacques and C. Vaziri, *Can. J. Chem.*, **49**, 1256 (1971); (e) K. von Bredow, H. Friebolin, and S. Kabuss, *Org. Magn. Resonance*, **2**, 43 (1970), and references therein; (f) K. von Bredow, H. Friebolin, and S. Kabuss in "Organic Chemistry: A Series of Monographs," Vol. 21, G. Chiurdoglu, Ed., Academic Press, New York, N. Y., 1971, p 51; (g) E. A. Noe and J. D. Roberts, *J. Amer. Chem. Soc.*, **93**, 7261 (1971); (h) E. Grunwald and E. Price, *ibid.*, **87**, 3139 (1965).
[5] A $^1/_8$ in. × 6 ft glass column of 10% SE-52 on 60–80 Chromosorb WAW was used; the column temperature was 80°.

Anet, et al. / Conformational Analysis of Cycloheptene Oxide