

75684-23-0; 13, 75626-30-1; 14, 75626-31-2; 15, 17302-56-6; 16, 17302-57-7; 18, 75626-32-3; 19, 75626-33-4; 20, 75626-34-5; 21, 75626-35-6; 22, 75684-24-1; 23, 75626-37-7; 24, 75684-25-2; 25, 75684-26-3; 26, 75626-37-8; 27, 75684-27-4; 28, 75684-28-5; 29,

75142-03-9; 29 S-oxide derivative, 75626-38-9; 30, 75626-39-0; toluquinone, 553-97-9; bis(*p*-chlorophenyl) disulfide, 1142-19-4; *p*-chlorobenzene-sulfonyl chloride, 933-01-7; *p*-aminophenol, 123-30-8; cyclopentadiene, 542-92-7.

Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data¹

Neil A. B. Gray, Christopher W. Crandell, James G. Nourse, Dennis H. Smith,*
Mary L. Dageforde, and Carl Djerassi*

Department of Chemistry, Stanford University, Stanford, California 94305

Received July 8, 1980

Computer programs for interpretation and prediction of ¹³C resonance spectra are described. These programs utilize a data base containing representations of the substructural environments of resonating nuclei together with their chemical shifts. These representations capture both molecular constitution and configuration, permitting for the first time in a computer program a comprehensive treatment of configurational stereochemical influences on ¹³C chemical shifts. Substructural features of an unknown structure are derived directly and automatically by ¹³C interpretive procedures. These features, together with additional structural information, are used to construct structural candidates for the unknown. ¹³C predictive procedures permit rank ordering of the candidates on the basis of agreement between predicted and observed ¹³C spectra. Applications of these programs to organic structure determination are illustrated through analyses of the structures of some diterpenes.

Early in the decade it was suggested that ¹³C spectroscopy was likely to prove the most important physical method of analysis in the 1970's.² ¹³C resonance data are indeed now routinely reported in papers on structure elucidation of natural products. Generally, however, the interpretation of a ¹³C spectrum is limited to deriving the number of CH₃'s, CH₂'s, and CH's from the multiplicities in the SFORD (single-frequency off-resonance decoupled) spectrum and determining the number of sp² carbon atoms from the gross chemical shifts. This is in spite of the fact that the chemical shift for a carbon nucleus is a sensitive probe of its stereochemical environment; observed shift values should reveal much about the bonding of individual atoms in an unknown structure. There is an obvious potential for automated systems that can help analyze ¹³C data in order to derive more of the implicit structural information. Such automated systems would aid both manual and computer-assisted explorations of structural possibilities for an unknown.

Programs for computer-assisted structure determination³⁻⁷ can make use of spectral data such as ¹³C resonances in two distinct ways. The first is in *interpretation* of data to obtain substructural information. These programs typically work by finding all ways of assembling or gen-

erating candidate structures for an unknown from substructural fragments (which may overlap in the case of GENOA⁴) whose presence has been inferred by interpretation of IR, proton resonance, and chemical data. Programs that could infer substructural fragments from ¹³C data would constitute a valuable adjunct to these standard sources of structural information. The second use is in *prediction* of spectra to evaluate the merit of each candidate. Once the structure-generating programs have created a set of candidate structures compatible with all given substructural constraints and, where appropriate, incorporating configurational stereochemistry,¹ the structural candidates can be ranked by determining some measure of how well each serves to rationalize observed spectral data. This approach has been applied previously in the context of mass spectral data.^{8,9} ¹³C spectral data can be used in similar fashion given suitable spectrum prediction and comparison functions.

In this paper we describe computer programs for inferring substructural information from ¹³C data and for ¹³C spectrum prediction together with ranking of candidate structures. The system that we have devised for these primary tasks has also proven of value in secondary applications such as aiding in the assignment of observed ¹³C resonances to the respective atoms of a known structure.

In contrast to the very extensive work on computer systems for processing mass spectral data,¹⁰ only a limited amount has been published on computerized analysis of ¹³C data. Pattern-recognition methods, for classifying ¹³C spectra according to the presence or absence of specific

(1) Part 34 of the series "Applications of Artificial Intelligence for Chemical Inference". For part 33 see J. G. Nourse, D. H. Smith, R. E. Carhart, and C. Djerassi, *J. Am. Chem. Soc.*, **102**, 6289 (1980).

(2) C. Djerassi, "¹³C NMR Spectroscopy", E. Breitmaier and W. Voelter, Eds., Verlag Chemie, Weinheim/Bergstr., Germany, 1974, preface.

(3) R. E. Carhart, D. H. Smith, H. Brown, and C. Djerassi, *J. Am. Chem. Soc.*, **97**, 5755 (1975).

(4) R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse, and C. Djerassi, *J. Org. Chem.*, in press.

(5) M. E. Munk, C. S. Sodano, R. L. McLean, and T. H. Haskell, *J. Am. Chem. Soc.*, **89**, 4158 (1967).

(6) S. Sasaki, Y. Kudo, S. Ochiai, and H. Abe, *Mikrochim. Acta*, 726 (1971).

(7) L. A. Gribov and M. E. Elyashberg, *CRC Crit. Rev. Anal. Chem.*, **8**, 110 (1979).

(8) N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary, *Anal. Chem.*, **52**, 1095 (1980).

(9) A. Lavanchy, T. Varkony, D. H. Smith, N. A. B. Gray, W. C. White, R. E. Carhart, B. G. Buchanan, and C. Djerassi, *Org. Mass Spectrom.*, **15**, 355 (1980).

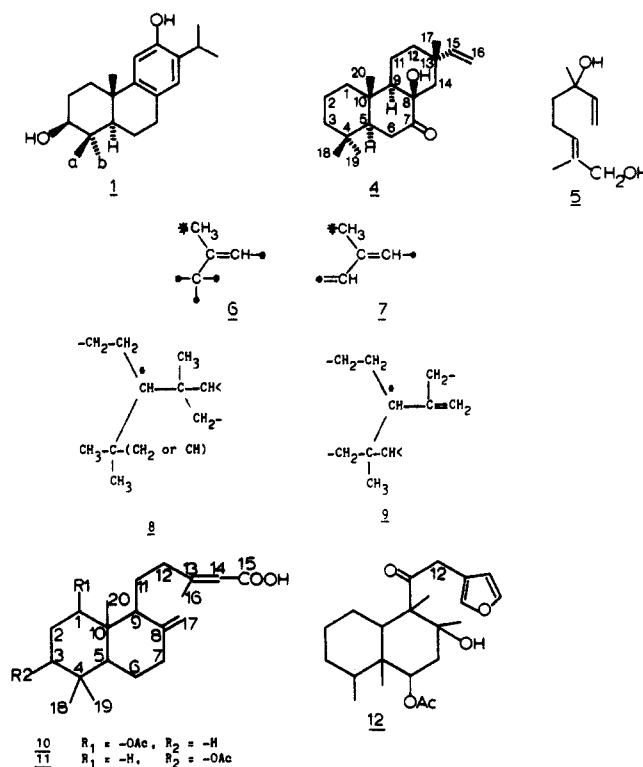
(10) G. M. Pesyna and F. W. McLafferty in "Determination of Organic Structures by Physical Methods", Vol. 6, F. C. Nachod, J. J. Zuckermann, and E. W. Randall, Eds., Academic Press, New York, 1976, p 91.

substructural features, have been reported¹¹ as have file-search methods for structure recognition.^{12,13} Clerc's file-search scheme¹² is now included as one component of a more elaborate ¹³C data base system.¹⁴ In addition to the overall spectrum matching option, this system provides a means for identifying structures associated with particular selections of resonance lines and a method for locating in the data base the chemical shifts of carbons in precisely defined environments.¹⁵ Spectrum-prediction programs have been reported for some limited classes of compounds,¹⁶⁻¹⁸ as has at least one class-specific scheme for the structural interpretation of ¹³C spectra.¹⁹

A variety of programs have been reported that in some way exploit a substructure-oriented approach that can serve both as a basis for spectrum prediction and, to a more limited extent, for spectrum interpretation. The "production rules" developed in our laboratory by Mitchell and Schwenzer^{20,21} were definitions of substructural environments and associated ranges for chemical shifts. This work was involved primarily with concepts of machine learning ("artificial intelligence"); the system was of limited practical utility for structure elucidation. Jezl and Dalrymple have reported a program using files of ¹³C spectra and fragment codes for chemical shift and environment matches.²² This system provides a number of options for spectrum analysis including overall spectrum matching, identification of substructures associated with particular resonance shifts, and prediction of resonances for atoms in partially specified chemical environments. An approach to complete structure elucidation, through analysis of retrieved substructural data, was described for the Jezl/Dalrymple system but was not implemented in a computer program. Bremser has developed a more comprehensive approach to processing ¹³C data that relies on a combination of overall spectrum matching, matching of "subspectra" and "substructures", and on exploiting a library of encoded atom environments.²³⁻²⁵ In Bremser's approach, the library of atom-centered codes and shifts is used as an aid to identifying substructural environments of those resonances not matched to any reference subspectrum/substructure combination. Possible applications to structure elucidation have been considered for the substructurally oriented DARC ¹³C data base.²⁶

The methods we present in the following sections are related to the work described above in that a data base relating substructural environments of resonating atoms to their chemical shifts is utilized in the computer processing of ¹³C NMR data to yield structural information

Chart I



for unknown compounds. However, our methods incorporate major extensions designed to overcome limitations of other systems, including the following: (1) unique and unambiguous ("canonical") characterization of substructural environments; (2) inclusion of configurational stereochemistry, essential for accurate characterization of structures and substructures; (3) an interpretive procedure which exploits the complementary structural information present in a ¹³C spectrum; (4) automation of all aspects of ¹³C data and structure processing.

Methods

(1) Representation of the Environment of a Resonating Nucleus. Our initial objective was to find a representation of the substructural environment of a resonating nucleus which is both amenable to computer processing and which captures enough of the structural factors responsible for an observed chemical shift to yield precise relationships between substructures and shifts. The various ¹³C processing systems mentioned in the introduction all employ some means for encoding substructural information. For several reasons, primarily the lack of consideration of stereochemistry, these other methods were unsuitable to meet our requirements. These reasons are discussed in another paper²⁷ which covers in detail our encoding scheme. Because we wished to capture the important influences of stereochemistry on chemical shifts, we began with the requirement that configurational stereochemistry be an integral part of our structure and substructure descriptions.

The problem of determining the number, nature, and representations of the different possible stereoisomers of a given structure has only recently been solved.²⁸ Historically, this was probably the reason why existing sub-

(11) C. L. Wilkins and T. R. Brunner, *Anal. Chem.*, **49**, 2136 (1977).

(12) R. Schwarzenbach, J. Meili, H. Konitzer, and J. T. Clerc, *Org. Magn. Reson.*, **8**, 11 (1976).

(13) V. Mlynarik, M. Vida, and V. Kello, *Anal. Chim. Acta*, **122**, 47 (1980).

(14) D. L. Dalrymple, C. L. Wilkins, G. W. A. Milne, and S. R. Heller, *Org. Magn. Reson.*, **11**, 535 (1978).

(15) J. Zupan, S. R. Heller, G. W. A. Milne, and J. A. Miller, *Anal. Chim. Acta*, **103**, 141 (1978).

(16) J. T. Clerc and H. Sommerauer, *Anal. Chim. Acta*, **95**, 33 (1977).

(17) D. H. Smith and P. C. Jurs, *J. Am. Chem. Soc.*, **100**, 3316 (1978).

(18) H. L. Surprenant and C. N. Reilley, *ACS Symp. Ser.*, **No. 54**, 77 (1977).

(19) R. E. Carhart and C. Djerassi, *J. Chem. Soc., Perkin Trans. 2*, 1753 (1973).

(20) T. M. Mitchell and G. M. Schwenzer, *Org. Magn. Reson.*, **11**, 378 (1978).

(21) G. M. Schwenzer and T. M. Mitchell, *ACS Symp. Ser.*, **No. 54**, 58 (1977).

(22) B. A. Jezl and D. L. Dalrymple, *Anal. Chem.*, **47**, 203 (1975).

(23) W. Bremser, M. Klier, and E. Meyer, *Org. Magn. Reson.*, **7**, 97 (1975).

(24) W. Bremser, *Z. Anal. Chem.*, **286**, 1 (1977).

(25) W. Bremser, *Anal. Chim. Acta*, **103**, 355 (1978).

(26) J. E. Dubois and J. C. Bonnet, *Anal. Chim. Acta*, **112**, 245 (1979).

(27) N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith, and C. Djerassi, *Org. Magn. Reson.*, in press.

(28) (a) J. G. Nourse, *J. Am. Chem. Soc.*, **101**, 1210 (1979); (b) J. G. Nourse, R. E. Carhart, D. H. Smith, and C. Djerassi, *ibid.*, **101**, 1216 (1979).

structurally oriented methods for processing ^{13}C data have either ignored stereochemical factors or have employed ad hoc formalisms for representing certain types of stereochemistry (e.g., cis/trans isomers on a double bond). This earlier omission of stereochemical detail is unfortunate because stereochemical factors, e.g., steric compression, play a major role in determining the magnetic environments experienced by the nuclei of atoms in a molecule and, consequently, their associated chemical shifts. Inclusion of configurational stereochemistry allows a substructure code to capture obvious important distinctions such as the cis/trans nature of substituents on both ring systems and double bonds and the differences between diastereomers of structures with chiral centers. We also need to represent more subtle features, such as the difference in the environments of diastereotopic substituents (e.g., methyl groups) whose atoms often display dramatically different chemical shifts. For example, the diastereotopic methyl groups a and b in structure 1²⁹ (see Chart I) resonate at 15.2 and 27.9 ppm, respectively.

The scheme we devised for encoding the substructural environment of an atom is described in detail elsewhere.²⁷ Briefly, a compact code, derived automatically, is used to represent the substructural environment of a resonating nucleus. A special algorithm adapted from earlier work on stereochemistry²⁸ is used to derive a standard form for a view of the stereochemical structure as seen from the resonating atom. Thus, the code captures molecular constitution and configuration. The standard form of the code is unique and unambiguous so that the same substructure in a variety of chemical structures will receive the same code.

The encoding scheme represents the *configurational stereochemical environment* of a resonating atom out to a four-bond radius about the atom, thereby including effects of δ substituents. Except for some special cases involving π systems,²⁷ the derived code can be viewed as a "shell structure". This shell structure represents the relative levels of importance of substituent groups on an atom's chemical shift.

Illustrations of the extent to which an atom's environment is captured by the encoding scheme are given in Figure 1. In structures 2 and 3 the resonating atoms are designated with an asterisk. The substructure captured in the code at each shell level is shown below the structure. The chiral centers in the structures are represented by stereochemical descriptors in the codes and indicated schematically in Figure 1.

Stereochemical information is encoded only at that shell level where a sufficient portion of the substructural environment has been included to allow stereochemical distinction. Thus, the cis/trans nature of a disubstituted alkene CH=CH is included at the "zero" shell level (as illustrated for 2, Figure 1) whereas the substructure code generated for a chiral center in a long chain will not include a stereochemical subcode if that center is sufficiently isolated from other stereocenters, i.e., no other stereocenters within the four-bond environment of the chiral center. The stereochemical code represents only relative and not absolute configurations since the normal ^{13}C resonance experiment does not yield data that could make distinctions about absolute stereochemistry. The stereochemical designations can be kept separate from the constitutional code. This allows the stereocode to be ignored if the data base of substructure codes is used in applications involving only structural topology (constitution).

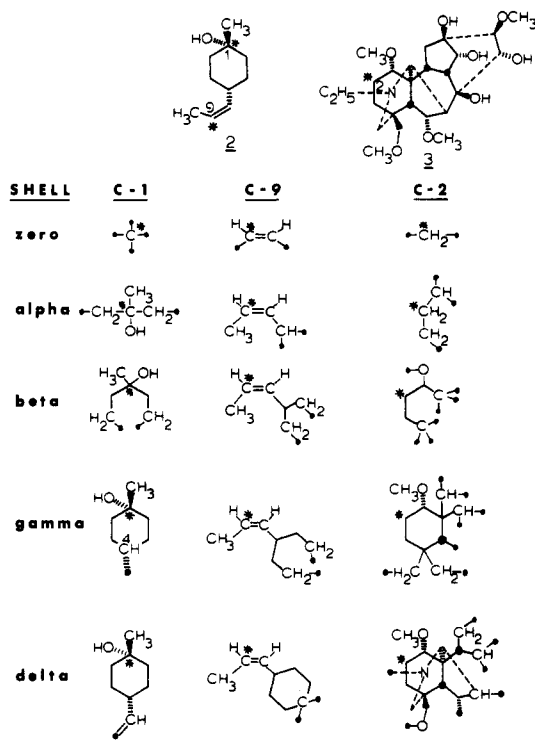


Figure 1. Illustrations of environments captured by the substructural coding scheme. The resonating atom is marked with an asterisk. In the shell representations of substructures in this figure, the configurational stereochemistry used by the program is described pictorially. Bonds with an unspecified terminus ("free valences") are bonds to nonhydrogen atoms in the next shell. If these free valences are sufficient to identify the chiral nature of the carbon atom, they are given specified orientations, for example, the free valence at C-4 at the γ shell for 2.

Conformational factors influence chemical shifts. Our coding scheme does not capture information on molecular conformations other than implicit specification of conformation resulting from designation of configurations, e.g., at ring junctures of rigid ring systems. In such instances, most substructures common to a family of related molecules will be found in the same conformational environment.

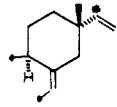
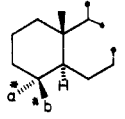
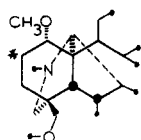
Our substructure codes with both molecular constitution and configuration capture most of the important influences on ^{13}C chemical shifts as evidenced by the very narrow ranges of shifts observed for the same substructure code in a wide variety of molecules.²⁷ Our representation has proven adequate for the development of a data base to be used in conjunction with ^{13}C spectrum interpretation and prediction. Data on selected substructures with their associated chemical shifts are presented in Table I to give some indication of the contents of the data base and the precision of chemical shifts.

(2) Data Base and Its Integrity. The successful application of either the interpretation or prediction programs is *critically dependent on the quality of the data base of reference substructures and spectral assignments*. Erroneous assignments for substructures in the data base can seriously impede both the interpretation and prediction processes. We have built our data base from structures and assigned ^{13}C spectra available in the literature. We have taken special precautions to ensure the integrity of the growing data base.

Systems for maintaining the integrity of the data base are described in more detail elsewhere.²⁷ Basically, the initial entries in the data base used published data in which chemical shift assignments were based on detailed evi-

(29) E. Wenkert, G. V. Baddeley, I. R. Burfitt, and L. N. Moreno, *Org. Magn. Reson.*, 11, 337 (1978).

Table I. Selected Substructures from the Data Base, Indicating the Precision of ^{13}C Chemical Shifts Obtained with the Four-Bond Radius Representation of Constitution and Configuration

| substructure | chemical shifts for starred atoms, ppm | |
|-----------------------------------------------------------------------------------|----------------------------------------|-------------------------------------|
|  | min | 149.1 |
| | max | 150.7 |
| | mean | 150.1 |
| | std dev | 0.4 |
| | 15 examples | |
|  | min | 33.0 ^a 21.0 ^b |
| | max | 34.5 22.5 |
| | mean | 33.5 21.6 |
| | std dev | 0.2 0.3 |
| | 57 examples | |
|  | min | 25.2 |
| | max | 26.5 |
| | mean | 26.0 |
| | std dev | 0.4 |
| | 16 examples | |

^a Values in this column are for atom a. ^b Values in this column are for atom b.

dence. Published data on other compounds, where assignments are based simply on chemical shift trends, are then verified against the more reliable data as the new compounds are entered.

Two factors serve to broaden chemical shift ranges observed for a given substructure. The first is solvent effects. Most of the data in the data base were obtained from spectra recorded by using CDCl_3 as a solvent. However, many spectra available in the literature were recorded in different solvents. The second factor is conformational effects which are not captured implicitly by the coding scheme. We accept the possibility of broadened shift ranges from these factors and take this into account in the programs for interpretation and prediction.

The data base consists of substructural codes with their associated chemical shift ranges. In effect, the data base is a compact computer representation of data such as those described in Table I. The interpretation and prediction programs access and utilize the compact representations in their computations. However, an investigator using the programs does not need to understand the details of the substructural codes; teletype-oriented drawings of the substructural environments represented by the codes are available on command.³⁰

The current data base includes data on monoterpenes,^{31,32} sesquiterpenes,³² diterpenes,³³ sesterpenes,³⁴ triterpenes,^{29,35} and steroids³⁶ along with a variety of

standards.³⁷ A file of spectra of diterpene alkaloids is also being constructed to allow more general testing of the utility of these programs.³⁸ The data base currently contains 10350 distinct substructure/shift pairs (7691 unique substructural codes) derived from a set of slightly less than 700 structures.

(3) Method for Spectrum Prediction. As described at the beginning of the paper, prediction of spectral properties plays an important role in our techniques for computer-assisted structure elucidation. Given the data base, which defines substructural environments and associated chemical shifts, a method for spectrum prediction is intuitively obvious. One merely treats the data base as a very large correlation table. Each candidate structure is processed by generating substructural environment codes for each of the carbon atoms in the structure by utilizing the same program used for generation of substructural codes during compilation of the data base. The latter is used by looking up the codes and retrieving the chemical shift range associated with each code. The look-up procedure for finding codes has been made quite efficient by using standard "hashing" schemes to provide the location into the data base, which is organized as a random-access file.

As illustrated in Table I, the data base includes information on the minimum, mean, and maximum shifts observed for a given substructure together with additional data such as the standard deviation of the shifts and the number of previously observed instances of the substructural environment. These data are used later to provide a measure of how reliable a particular prediction is likely to be.

Typically, the set of substructures included in the data base proves to be partially inadequate in that complete codes matching the environments of atoms in a candidate structure will not be found. For example, the current data base has few substructures involving sulfur or halogens. In such cases, the spectrum-prediction functions try successively to match codes for three-, two-, or even one-shell environments. (The encoding scheme has been designed so that substructures with the same environment out to an n -bond radius will have precisely the same code out to the n th shell.²⁷) If codes are found to match successfully only at an inner shell the shift ranges associated with substructures possessing the same code out to that shell are retrieved. Such less specific substructural environments are associated with increasingly wide ranges of chemical shift as exemplified below.

The result of these prediction procedures is a "fuzzy" predicted spectrum consisting of the set of shift ranges retrieved for each carbon atom in the candidate structure. As an illustration of the appearance of such a fuzzy spectrum, the chemical shift ranges retrieved for the carbon atoms of "compactone"³⁹ (4) are presented in Table II. The ranges in predicted shift values vary widely for different atoms. Some atoms, e.g., C-3, are in fairly standard substructural environments corresponding to three/four shell substructures in the library; for such atoms relatively precise predictions can be obtained. Other atoms such as C-7 have few or no prototypes in the data base, and predictions must be made on the basis of one-bond matching environments. Inevitably such predictions lead to broad ranges. In still other cases, such as C-17, pre-

(30) R. E. Carhart, *J. Chem. Inf. Comput. Sci.*, **16**, 82 (1976).

(31) F. Bohlmann, R. Zeisberg, and E. Klein, *Org. Magn. Reson.*, **7**, 426 (1975).

(32) F. W. Wehrli and T. Nishida, *Fortschr. Chem. Org. Naturst.*, **36**, 1 (1978).

(33) (a) J. Fayos, M. Martinez-Ripoll, M. Paternostro, F. Piozzi, B. Rodriguez, and G. Savona, *J. Org. Chem.*, **44**, 4992 (1979), and references cited therein; (b) I. Wahlberg, K. Nordfors, M. Curvall, T. Nishida, and C. R. Enzell, *Acta Chem. Scand., Ser. B*, **33**, 437 (1979), and references cited therein; (c) A. Arnone, R. Mondelli, and J. St Pyrek, *Org. Magn. Reson.*, **12**, 429 (1979); (d) W. Herz, S. V. Govindan, and J. F. Blount, *J. Org. Chem.*, **44**, 2999 (1979), and references cited therein; (e) E. Wenkert, P. Ceccherelli, M. S. Raju, J. Polonsky, and M. Tingoli, *ibid.*, **44**, 146 (1979); (f) B. Delmond, B. Papillaud, J. Valade, M. Petraud, and B. Barbe, *Org. Magn. Reson.*, **12**, 209 (1979); (g) K. C. Joshi, R. K. Bansal, T. Sharma, R. D. H. Murray, I. T. Forbes, A. F. Cameron, and A. Maltz, *Tetrahedron*, **35**, 1449 (1979).

(34) G. Cimino, S. DeStefano, L. Minale, and E. Trivellone, *J. Chem. Soc., Perkin Trans. 1*, 1587 (1977).

(35) G. S. Ricca, B. Danielli, G. Palmisano, H. Duddeck, and M. H. A. Elgamal, *Org. Magn. Reson.*, **11**, 163 (1978).

(36) J. W. Blunt and J. B. Stothers, *Org. Magn. Reson.*, **99**, 439 (1977).

(37) P. A. Couperus, A. D. H. Clague, and J. P. C. M. van Dongen, *Org. Magn. Reson.*, **11**, 590 (1978), and references cited therein.

(38) J. Finer-Moore, private communication.

(39) A. C. Pinto, *Phytochemistry*, **18**, 2036 (1979).

Table II. Predicted Spectral Data^a for "Compactone" (4)^b

| atom | SHELL | RESMIN | RESMAX | RESAVG | RESONANCES | atom | SHELL | RESMIN | RESMAX | RESAVG | RESONANCES |
|------|-------|--------|--------|--------|------------|------|-------|--------|--------|--------|------------|
| 1 | 3 | 37.8 | 42.0 | 39.8 | 73 | 11 | 2 | 14.1 | 23.1 | 19.2 | 134 |
| 2 | 3 | 18.0 | 23.5 | 18.7 | 72 | 12 | 3 | 35.7 | 37.5 | 36.6 | 2 |
| 3 | 3 | 40.9 | 42.5 | 41.9 | 93 | 13 | 2 | 33.6 | 37.9 | 35.3 | 7 |
| 4 | 2 | 32.5 | 33.9 | 33.2 | 93 | 14 | 1 | 39.8 | 58.0 | 47.6 | 42 |
| 5 | 1 | 37.5 | 62.5 | 53.1 | 371 | 15 | 3 | 146.2 | 150.2 | 148.3 | 7 |
| 6 | 2 | 29.1 | 37.8 | 35.0 | 4 | 16 | 3 | 108.9 | 111.9 | 110.4 | 7 |
| 7 | 1 | 186.1 | 224.5 | 213.0 | 37 | 17 | 3 | 23.0 | 32.3 | 27.1 | 7 |
| 8 | 1 | 81.8 | 81.8 | 81.8 | 1 | 18 | 4 | 32.3 | 32.8 | 32.6 | 3 |
| 9 | 1 | 37.5 | 62.5 | 53.1 | 371 | 19 | 4 | 20.7 | 20.8 | 20.7 | 3 |
| 10 | 2 | 32.7 | 40.3 | 37.8 | 92 | 20 | 3 | 11.3 | 18.1 | 15.6 | 90 |

^a "SHELL" specifies the degree of match between the substructural environment of an atom in compactone and the substructure used as a basis for prediction. The RESMIN and RESMAX values give the minimum and maximum shift values associated, in the data base, with an atom in the equivalent substructural environment as that observed in compactone. RESAVG is the average shift associated with such substructures. RESONANCES indicates the number of examples of this substructure observed previously among the structures used for the creation of the data base.

dictions may be derived from prototype substructures in two different configurational forms, and the resulting range may again be large (the shell-four environment is needed to define the configuration for C-17).

While spectrum prediction based on the above procedure is conceptually straightforward, comparative evaluation of the predicted spectra for several alternative candidate structures is more complex. First, some measure of similarity between a fuzzy, predicted spectrum and an observed spectrum must be calculated. Second, the alternative structures must be ranked on the basis of these similarity measures.

Complexities arise in the computation of a fair similarity measure when matching the predicted and observed spectra. It is inappropriate for the similarity measure to give a poor score to a candidate structure for which the resonance shift predictions were based on poor, e.g., one- or two-shell, matching of the substructure codes. If a structure has several poor matches during prediction, the fault is due to limitations of the data base and *not* to the structure itself. Therefore, the similarity measure must be conservative so that there is no discrimination against candidates with substructures poorly represented in the data base. *Only structures which result in dissimilar predicted and observed spectra based on adequate substructure representation in the data base should be given significantly poorer scores.*

The spectrum matching process involves two steps: (1) Ascribe each predicted fuzzy resonance to an observed resonance of the correct multiplicity. In the current program the mean values for the predicted ranges are used as a basis for sorting the predicted resonances in order of chemical shift and placing them into correspondence with observed resonances similarly sorted. (2) Compute a score for each resonance based on the observed and predicted chemical shifts. The results presented below are based on a simple *dissimilarity* measure in which the mismatch between predicted and observed resonances is scored as being proportional to the square of the difference between the predicted mean and observed shift, weighted by the shell level from which the predicted range was derived, as shown in eq 1, where S_i is the shell-level of the substructure

$$\text{resonance-score}_i = S_i * (M_i - O_i)^2 / 8 \quad (1)$$

code used to predict the resonance of atom i , M_i is the mean value of the shift range predicted for the resonance of atom i , and O_i is the observed shift associated with the resonance predicted for atom i .

The overall score for matching of the complete spectrum is simply the sum of the scores for the individual resonances (eq 2). Because the dissimilarity between the spectra is being computed, the higher the overall score, the poorer the match.

$$\text{candidate score} = \sum_{i=1}^n \text{resonance-score}_i \quad (2)$$

The weighting of the resonance score according to the shell level of the substructure code used for prediction (eq 1) is found in practice to bias slightly this scoring toward structures whose substructure codes are poorly represented in the data base (small S_i values). The quadratic dependence of the scores on the "error" in prediction (eq 1) has generally provided a good degree of discrimination between similar structures represented by similar detailed three- to four-shell substructure codes. The quantity eight in the denominator of eq 1 is an attenuation factor. More elaborate functions that provide better discrimination and yet are still tolerant of limitations of the data base are

currently under development.⁴⁰

(4) **Method for Spectrum Interpretation.** The same data base can be used in interpretation of ¹³C NMR spectra. Here the procedure is again, in principle, intuitively obvious. Given the spectrum of an unknown compound, for each observed resonance retrieve from the data base all substructures which display similar resonances. Then, attempt to piece together the complete structure of the unknown on the basis of examination of the retrieved substructures. In practice, however, for molecules for moderate complexity the number of substructures retrieved and the vast number of ways in which they could be pieced together defy analysis without some computational assistance. (Obviously, if one restricts the problem to new structures which are known to be closely related to a set of previously assigned spectrum/structure pairs, there is a much greater chance of solving the problem manually.) For example, given a C₂₀ molecule with several rings and heteroatoms and reasonable tolerances on matching of shifts, an initial check against the data base can result in (2–3) × 10³ unique substructures consistent with the molecular formula and matched to individual resonances. Obviously, direct intercomparison of so many substructures is impractical.

A further complication of the procedure is that we cannot assume that the data base contains detailed, four-bond environments for all atoms in an unknown. In fact, the limited size of our data base (and all other data bases for the next several years) makes it highly probable that *no* atom will have its environment described exactly to a four-bond radius unless closely related compounds have been entered in the data base. The immense number of possible substructure environments when several heteroatoms and stereochemical configurations are included forces us to restrict the interpretation procedure to consider substructural consistencies to a one- or two-bond radius where there is a high probability that *every* atom will have its environment represented to that shell level.

Consider an individual ¹³C resonance of given multiplicity and chemical shift. This single resonance generally permits numerous substructural interpretations. As an example, 12 different two-bond substructures were associated with a doublet resonance in the range 35.3–36.3 ppm. These alternative structures are shown in Figure 2. It would, in fact, be possible to use these 12 different substructures as alternative features to be built into candidate structures by the GENOA program.⁴ However, the variety of structural forms is, even in this simple case, sufficiently great to reveal that such a procedure would not effectively constrain the structure-generation process. Individual resonance lines allow too much ambiguity in interpretation.

If, however, the complete ¹³C spectrum is available, then much of the ambiguity in the interpretation of each individual resonance can be eliminated. Our interpretation procedure exploits the redundant structural information present in a complete spectrum by (1) identifying consistencies in the substructures associated with individual resonances and (2) intercomparing these consistent features among the complete set of resonances. Substructures can only be meaningfully associated with individual resonance lines if they are also consistent with the substructural interpretations found for other resonances in the spectrum. By performing these consistency checks in the computer, we can dramatically reduce the number of alternative interpretations for each resonance.

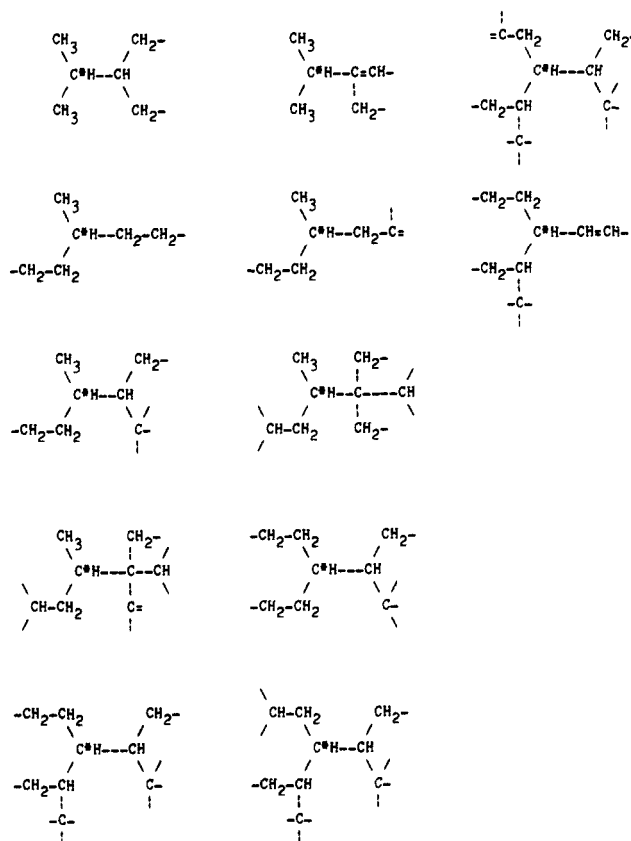


Figure 2. Twelve different substructures retrieved from a subset of the data base for a doublet resonance with a ¹³C chemical shift of 35.8 ± 0.5 ppm. The resonating atom is marked with an asterisk.

As will be illustrated below, we accomplish partial interpretation of a ¹³C spectrum using an iterative procedure to perform the consistency checks. First, the data base is searched to find those substructures that may be associated with each individual resonance considered in isolation. Subsequently, the substructures associated with the different resonance lines are intercompared and any found to be inconsistent eliminated. The intercomparison and elimination procedures are repeated until no further increase in specificity can be attained. For many structural problems this procedure results in detailed specification of the local environments of enough carbon atoms that generation of structural candidates is possible from the ¹³C interpretations themselves. (Currently, the structure-generating algorithms used^{3,4} work purely in terms of constitutional structure with the generation of configurational stereoisomers being accomplished subsequently by means of a separate program.^{1,28} Because stereochemistry is not currently used in initial structure generation, the interpretive procedures ignore all configurational information in the reference substructure codes.) Often, additional structural inferences from other spectroscopic techniques are also utilized simply because ¹³C NMR data alone are insufficient to yield the correct structure, and no chemist would attempt to solve a large structure solely on the basis of ¹³C data. However, as we illustrate in subsequent examples, the interpretive procedure *alone* can yield a surprising amount of structural information.

The interpretive procedures are in the form of an interactive computer program which allows several types of constraints to be supplied by the investigator, including matching tolerances for chemical shift values used to retrieve substructures from the data base and the shell levels used to specify to the program where substructural in-

(40) C. W. Crandell and N. A. B. Gray, to be submitted for publication in *J. Chem. Inf. Comput. Sci.*

Table III. ¹³C Spectrum of C₁₀H₁₈O₂ Monoterpenediol⁴¹

| atom | shift, δ | multiplicity | atom type |
|------|----------|--------------|------------------|
| 1 | 13.7 | q | CH ₃ |
| 2 | 27.6 | q | CH ₃ |
| 3 | 22.4 | t | CH ₂ |
| 4 | 41.8 | t | CH ₂ |
| 5 | 68.3 | t | CH ₂ |
| 6 | 111.8 | t | =CH ₂ |
| 7 | 125.7 | d | =CH |
| 8 | 144.9 | d | =CH |
| 9 | 73.3 | s | >C< |
| 10 | 134.9 | s | =C< |
| 11 | | | OH |
| 12 | | | OH |

terpretations must match. Matching windows of different shift widths can be defined for each observed resonance. Typically one might use a narrower shift range, e.g., ± 0.75 ppm, for substructures such as alkyl methyl groups which are well represented in the data base, while a wide range, e.g., ± 5 ppm, might be used for less common functionalities. These widths can be adjusted individually at the discretion of the investigator during the procedure in order to make the matching more tolerant for "unusual" observed resonance shifts, for example, to allow for conformational or solvent effects mentioned previously. The interpretation program can be used with a requirement for one-, two-, three-, or even four-bond substructural matches, and, again, different matching requirements can be imposed for each resonance line to allow for the fact that certain substructures may be well represented in the data base, while novel substructures such as those involving unusual heteroatoms may have few or no representatives.

We illustrate the essential aspects of the ¹³C spectral interpretation program through a description of a simple structural problem. The ¹³C data are taken from a recent investigation of the structure of a monoterpenediol isolated from Greek tobacco.⁴¹ The interpretation functions were given the molecular formula, C₁₀H₁₈O₂, and the complete ¹³C spectrum with line multiplicities from off-resonance decoupling as summarized in Table III. In this simple case, the program's preliminary analysis determines all atom types unambiguously from the molecular formula, line multiplicities, and chemical shift values (second column, Table III).

The resonance data were checked in more detail against the data base. The program was requested to retrieve as candidate interpretations for each resonance line those substructures with resonances of the appropriate multiplicity within 2.5 ppm of each observed resonance. A further restriction on retrieved substructures was that each had to incorporate just those atom types inferred as present (Table III) in the unknown molecule. Of the 1 × 10⁴ substructures in our file, only 238 passed these tests.

At this point the program examined the substructures associated with each resonance line to find common features. Results of this examination are shown in Figure 3 together with a random selection of some of the substructures from which the common features (out to the first shell) were detected. Note that the program has already begun to describe the substructural environment of the carbons at lines 4, 5, and 8–10 (Figure 3).

Examination of the common features in Figure 3 reveals several additional constraints implied by the substructures. For example, although both CH₂=C< and CH₂=CH substructures may be associated with the triplet resonance

| Resonance Line | Common substructure | Some of the alternative prototypes. |
|----------------------------------------------|--------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 (13.7, Q) (59 possible substructures) | -CH ₃ | $\begin{array}{c} \text{C}^*\text{H}_3-\text{CH}_2-\text{CH}_2- \\ \text{C}^*\text{H}_3-\text{CH}=\text{CH}- \\ \text{C}^*\text{H}_3 \\ \diagdown \\ \text{C}=\text{CH}- \\ \diagup \\ -\text{CH}_2 \end{array}$ $\begin{array}{c} \text{C}^*\text{H}_3-\text{CH}_2-\text{CH}- \\ \text{C}^*\text{H}_3 \\ \diagdown \\ \text{C}=\text{CH}- \\ \diagup \\ -\text{C}- \\ \diagdown \\ \text{C}=\text{C} \end{array}$ |
| 2 (27.6, Q) (28 possible substructures) | -CH ₃ | $\begin{array}{c} \text{C}^*\text{H}_3-\text{C}- \\ \\ \text{C}^*\text{H}_3-\text{C}=\text{C} \end{array}$ |
| 3 (22.4, T) (38 possible substructures) | -CH ₂ - | $\begin{array}{c} \text{CH}_3-\text{C}^*\text{H}_2-\text{CH}_2- \\ -\text{CH}_2-\text{C}^*\text{H}_2-\text{CH}_2- \\ -\text{CH}_2-\text{C}^*\text{H}_2-\text{CH}=\text{C} \end{array}$ $\text{CH}_3-\text{C}^*\text{H}_2-\text{CH}=\text{C} <$ |
| 4 (41.8, T) (14 possible substructures) | -C [*] H ₂ -CH ₂ - | $\begin{array}{c} -\text{CH}_2-\text{C}^*\text{H}_2-\text{C} \\ \\ -\text{CH}_2-\text{C}^*\text{H}_2-\text{C} \\ \diagdown \\ \text{C} \end{array}$ |
| 5 (68.3, T) (2 possible substructures) | HO-C [*] H ₂ -C=CH- | $\begin{array}{c} \text{HO}-\text{C}^*\text{H}_2-\text{C}=\text{CH}- \\ \\ \text{CH}_3 \\ \text{HO}-\text{C}^*\text{H}_2-\text{C}=\text{CH}- \\ \\ -\text{CH}_2 \end{array}$ |
| 6 (111.8, T) (33 possible substructures) | C [*] H ₂ = | $\begin{array}{c} \text{C}^*\text{H}_2=\text{CH}- \\ \text{C}^*\text{H}_2=\text{C} < \end{array}$ |
| 7 (125.7, D) (27 possible substructures) | -C [*] H= | $\begin{array}{c} -\text{C}^*\text{H}=\text{CH}- \\ -\text{C}^*\text{H}=\text{C} < \end{array}$ |
| 8 (144.9, D) (6 possible substructures) | $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_2=\text{C}^*\text{H}-\text{C}-\text{CH}_2- \\ \end{array}$ | $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_2=\text{C}^*\text{H}-\text{C}-\text{CH}_2- \\ \\ \text{CH}_2- \\ \\ \text{CH}_3 \\ \\ \text{CH}_2=\text{C}^*\text{H}-\text{C}-\text{CH}_2- \\ \\ \text{CH}=\text{C} \\ \\ \text{CH}_3 \\ \\ \text{CH}_2=\text{C}^*\text{H}-\text{C}-\text{CH}_2- \\ \\ \text{OH} \end{array}$ |
| 9 (73.3, S) (10 possible substructures) | $\begin{array}{c} \text{CH}_2- \\ \\ -\text{C}^*- \\ \\ \text{OH} \end{array}$ | $\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{C}^*-\text{CH}_2- \\ \\ \text{OH} \end{array}$ $\begin{array}{c} \text{CH}_3 \\ \\ -\text{CH}_2-\text{C}^*-\text{CH}_2- \\ \\ \text{OH} \end{array}$ $\begin{array}{c} \text{CH}_2- \\ \\ \text{CH}_3-\text{C}^*-\text{CH} \\ \\ \text{OH} \end{array}$ $\begin{array}{c} \text{CH}_2- \\ \\ -\text{CH}_2-\text{C}^*-\text{C} \\ \\ \text{OH} \end{array}$ |
| 10 (134.9, S) (21 possible substructures) | $\begin{array}{c} \text{CH}_3-\text{C}^*=\text{CH}- \\ \end{array}$ | $\begin{array}{c} \text{CH}_3-\text{C}^*=\text{CH}- \\ \\ \text{CH}_3 \end{array}$ $\begin{array}{c} \text{CH}_3-\text{C}^*=\text{CH}- \\ \\ \text{CH}_2- \end{array}$ |

Figure 3. Common substructural features identified from the alternative substructures retrieved for each of the resonance lines (±2.5 ppm) of the monoterpenediol spectrum (Table III).⁴¹

at 111.8 ppm (line 6), the substructures found for the doublet at 144.9 ppm (line 8) imply that only CH₂=CH substructures are appropriate. Other implications are less obvious. For example, associating the triplet at 22.4 ppm (line 3) with the substructure CH₂C^{*}H₂CH= (Figure 4) requires checking a long sequence of indirect implications. Such checks are automatically carried out in the program through a detailed investigation of possible bonds which

(41) D. Behr, I. Wahlberg, T. Nishida, and C. R. Enzell, *Acta Chem. Scand., Ser. B*, **32**, 228 (1978).

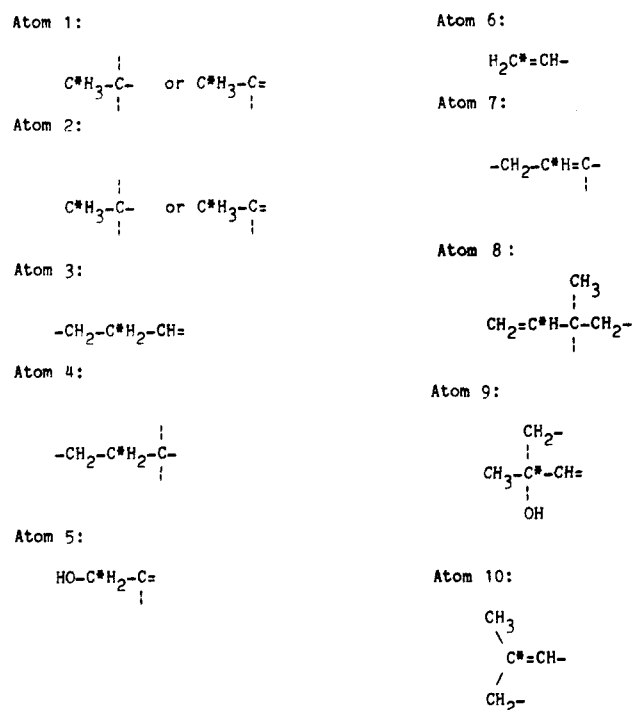


Figure 4. Characteristic substructural environments identified by intercomparison of substructures initially identified as possible for each of the individual resonance lines.

might be formed among individual atoms by using a connection matrix representation of the structure.⁴² The substructural constraints thus inferred for each atom are summarized in Figure 4.

Note that for every atom the program has derived a more detailed description of the environment than that available from considering the atoms in isolation (compare Figures 3 and 4). Thus, the program has made use of many constraints implied through consideration of the ¹³C spectrum in its entirety.

More information can be obtained, however. The implied bonding constraints used to derive environments in Figure 4 limit the allowed environments of each atom and can be used to eliminate many of the substructural prototypes retrieved during the initial pass through the data base. For example, because atom 6 is established to be part of a $\text{CH}_2=\text{CH}$ fragment (Figure 4), the program eliminates all substructures involving $\text{CH}_2=\text{C}<$ initially retrieved for line 6 (Figure 3). Performing these checks for all resonances reduces the number of possibly relevant substructures from 238 to 127. The program then performs a second complete iteration with the reduced set of possible substructures for the individual resonances. Additional consistent features are found in the sets of substructural prototypes that remain, resulting, in this example, in a new set of substructural environments which are summarized in Figure 5. Only one structure can be assembled from these substructures, 2,6-dimethyl-2,7-octadiene-1,6-diol (5), the structure originally assigned on the basis of a variety of spectral data.⁴¹ We verified the authors' assignment of the configuration of the C-2,3 double bond (*E*) by predicting spectra for both the *Z* and *E* isomers and comparing the predicted spectra with that observed. Relevant prototype substructures in the data base for this comparison were derived from the collections of ref 32 and 38.

In this example, no further iterative refinement of the substructural environments could be achieved without

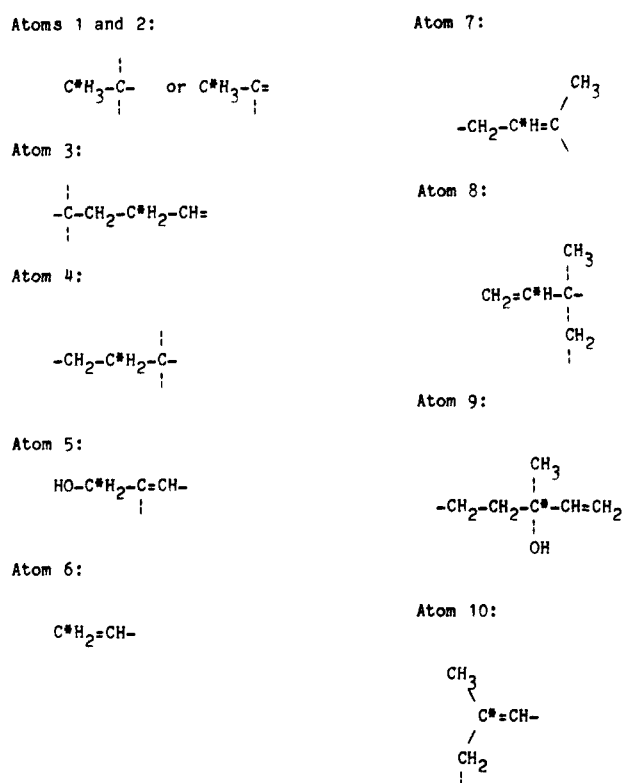


Figure 5. Final substructural environments derived for the atoms of the monoterpene diol 5.

specifying narrower resonance ranges or requiring more detailed checks on the forms of the substructures. For example, a requirement of an exact two-bond match for all atoms would, in fact, eliminate many of the prototype substructures still retained as valid alternative interpretations of individual resonance lines, allowing additional inferences. Thus, the quartet at 13.7 ppm (line 1, Figure 3) is associated with, among others, the substructures 6 and 7. Both of these could be eliminated by checking the complete two-bond environment of their methyl groups because the bonding constraints already identified for the $>\text{C}=\text{C}$ atom (line 10, Figure 3) in the unknown structure prohibit single bonds to quaternary alkyl carbons and to conjugated $\text{CH}=\text{C}$ atoms.

In general, however, for larger structural problems of biological significance, several iterations yield successive refinements on bonding constraints, and thus, relevant substructures initially retrieved for each resonance. The iterations are continued until no further refinement is achieved.

Results

In this section we present two examples which illustrate the performance of the spectral interpretation and prediction programs. These examples, involving polyfunctional diterpenoid structures previously identified by other workers on the basis of examination of a variety of spectral data, are representative of a large number of cases we have examined and demonstrate that considerably more structural information can be derived from a ¹³C spectrum than is commonly realized by chemists.

(1) **Interpretation of Spectra.** This illustration of the ¹³C spectrum interpretation program uses published data on the identification of a labdanoid diterpene isolated from the autumnal leaves of *Metasequoia glyptostroboides*.⁴³

(42) N. A. B. Gray, *Anal. Chem.*, **47**, 2426 (1975).

(43) S. Braun and H. Breitenbach, *Tetrahedron*, **33**, 145 (1977).

Table IV. ^{13}C Spectrum of "Unknown" $\text{C}_{22}\text{H}_{34}\text{O}_4$ Used in the Test of Spectrum-Interpretation Functions⁴³

| shift, δ | mult ^a | shift, δ | mult ^a | shift, δ | mult ^a | shift, δ | mult ^a |
|--------------------|-------------------|--------------------|-------------------|--------------------|-------------------|--------------------|-------------------|
| 14.5 | q | 21.8 | t | 107.0 | t | 38.1 | s |
| 16.5 | q | 24.0 | t | 54.8 | d | 39.3 | s |
| 19.2 | q | 24.3 | t | 55.9 | d | 147.5 | s |
| 21.3 | q | 36.8 | t | 80.7 | d | 163.4 | s |
| 28.3 | q | 38.1 | t | 115.1 | d | 171.0 | s |
| | | 40.0 | t | | | 171.8 | s |

^a Multiplicity.

The ^{13}C spectrum of this $\text{C}_{22}\text{H}_{34}\text{O}_4$ compound is summarized in Table IV. This compound is similar to many of the structures used to create the current data base but is not itself included in the data base.

The molecular formula and spectral data were supplied to the interpretive program, and the iterative cross-checking sequence was initiated. The constraints imposed required that the substructures retrieved for methyl groups be those with resonances within ± 0.75 ppm of an observed quartet resonance, and other substructures had only to match within ± 1.5 ppm of an observed resonance of appropriate multiplicity. At this point in the processing we requested that substructures be consistent at shell two.

The iterative process converged, and no further constraints could be identified when 393 distinct substructures remained. Some of the substructures that remained were inappropriate on the basis of information available from other spectroscopic techniques. Thus, the quartet resonance at 14.5 ppm was associated with ethyl groups contained in a variety of larger substructural environments (as well as the "correct" CH_3C substructures). Because at least one of the methyl groups could potentially bond to a CH_2 , the sets of substructures identified as plausible for each of the five CH_2 triplet resonances in the range 21–38 ppm all include some $\text{CH}_2\text{*CH}_3$ in addition to other substructural environments. Although it was not done for this example illustration, the fact that the proton spectrum showed all the methyls resonating as singlets could be used to eliminate the inappropriate ethyl substructures. Use of this information would have allowed the iterative procedure to be continued further, yielding more precise descriptions of the environments of the methyl and methylene carbons. The interpretation program does provide limited facilities for the investigator to utilize substructural information derived from other sources; alternatively additional information can be used during structure generation with GENOA, as discussed below.

To some extent, the spectrum-interpretation program is handicapped by attempting to solve two distinct problems simultaneously. The program is not just deriving the structure but it is also effectively making a spectral assignment on the basis of shift data. Some assignments may be ambiguous, and this can inhibit the structure elucidation process. This is illustrated by consideration of the two methine resonances at 54.8 and 55.9 ppm (Table IV). The substructural constraints derived for these two resonances, in combination with other constraints for other resonances, implied the presence of substructures 8 and 9. But the two different resonances are equally readily assigned to either of these substructures. Because there is no way for the program to select which resonance to assign to which substructure, identification of the common features associated with these resonances is limited to specifying that they correspond to $>\text{CH}$'s bonded to CH_2 's and to one or more $>\text{C}<$'s.

The ^{13}C spectrum-interpretation functions are interfaced with the GENOA structure-generating program.⁴ GENOA was

given the substructural data derived for the compound, as summarized in Figure 6. The data in the "additional constraints inferred" column of Figure 6 represent information which cannot be expressed in precise substructural terms but which could be inferred from manual examination of the output of the interpretation program. This additional information was supplied to GENOA, by using the interactive interface to the program, in the form of substructural constraints which were then employed in the structure-generation procedure.

The authors' spectroscopic and chemical data, including similarities to compounds isolated previously,⁴³ pointed to the presence of a labdane skeleton. As a test of that hypothesis we assumed during generation of candidate structures with GENOA that the unknown incorporated one of several possible common diterpenoid skeletons, including not only the labdane skeleton but also the clerodane, pimarane, cassane, abietane, totarane, taxane, kaurane, beyerane, atisane, aconane, and gibbane skeletons.⁴⁴ These skeletal systems were used to define the initial range of structural forms. The substructures inferred from the ^{13}C data were then applied as constraints. Just two structural isomers, irrespective of stereochemistry, were obtained, 1-acetoxylabda-8(17),13-dien-15-oic acid (10) and 3-acetoxylabda-8(17),13-dien-15-oic acid (11). (The numbering system we used for the labdane skeleton differs from that in ref 43, the difference being simply an interchange of atoms 17 and 20.) If, during the interpretive procedure, we had chosen to eliminate the possibility of ethyl groups as discussed previously and continued the interpretive procedure further, then the more complete environment identified for the methyl associated with the 14.5-ppm quartet would have served to eliminate the 1-acetoxy substructure, thereby leading to a unique constitution 11.

Because configurational stereochemistry had been ignored during the interpretation and structure-generation procedures, there still remained the problem of determining the appropriate stereochemistry. If a standard labdane-type skeleton is assumed, then the configurations at atoms 5, 9, and 10 may be fixed. Since only relative and not absolute stereochemistry is considered, four stereoisomers are possible for each of 10 and 11. These differ in the cis/trans nature of the C-13,14 double bond and the axial/equatorial nature of the acetoxy substituent on ring A.

The eight stereoisomers were generated¹ and their ^{13}C spectra predicted. The resulting predictions were used as a basis for ranking the candidates. These ranking results established that the most likely structure was, as originally determined by Braun and Breitenbach,⁴³ the 3-acetoxy, C-13,14 trans isomer with a β acetoxy group. The second most likely candidate was the isomer with the 3-acetoxy group in an α configuration. The ranking procedure also served to determine the quality of the file, in terms of relevant substructures, for this "unknown". While most atoms had three-bond or even four-bond substructural environment matching codes in the file, a few such as C-5 and C-10 had at best two-bond substructural matches.

(2) Spectrum Prediction and Structure Ranking. The prediction and ranking procedures of our program were carried out with another labdanoid diterpene,⁴⁵ $\text{C}_{22}\text{H}_{32}\text{O}_5$, whose ^{13}C spectrum is shown in Table V. Attempted interpretation of the spectrum by using the same procedures as for the labdane from *Metasequoia glyp-*

(44) J. R. Hanson in "Chemistry of Terpenes and Terpenoids", A. A. Newman, Ed., Academic Press, New York, 1972.

(45) O. Prakash, D. S. Bhakuni, R. S. Kapil, G. S. R. Subbo Rao, and B. Ravindranath, *J. Chem. Soc., Perkin Trans. 1*, 1305 (1979).

| Resonance | Substructure passed to GENOA. | Additional constraints inferred. |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| 14.5 Q | -CH ₃ | Bonded to -CH ₂ - or >C<. |
| 16.5 Q | -CH ₃ | Bonded to >C< or >C=. |
| 19.2 Q | -CH ₃ | |
| 21.3 Q | -CH ₃ | |
| 28.3 Q | $\begin{array}{c} \text{CH}_3 \\ \\ -\text{C}- \\ \end{array}$ | |
| 21.8 T | -C [*] H ₂ -CH ₂ - | Each has identical bonding constraints: 0-1 CH ₃ ^s 1-2 CH ₂ ^s 0-1 CH ₃ ^s |
| 24.0 T | -C [*] H ₂ -CH ₂ - | |
| 24.3 T | -C [*] H ₂ -CH ₂ - | |
| 36.8 T | -CH ₂ - | |
| 38.1 T | -CH ₂ - | |
| 40.0 T | -C [*] H ₂ -CH ₂ - | Bonded to either >C< or >C=. |
| 107.0 T | $\begin{array}{c} \text{CH}_2- \\ / \\ \text{C}^*\text{H}_2=\text{C} \\ \backslash \\ -\text{CH}- \end{array}$ | |
| 54.8 D | $\begin{array}{c} \text{CH}_2- \\ \\ -\text{C}^*\text{H} \\ \\ -\text{C}- \\ \end{array}$ | Also bonded to either >C< or >C=. |
| 55.9 D | $\begin{array}{c} \text{CH}_2- \\ \\ -\text{C}^*\text{H} \\ \\ -\text{C}- \\ \end{array}$ | Also bonded to either >C< or >C=. |
| 80.7 D | $\begin{array}{c} -\text{CH}_2 \\ \\ \text{C}^*\text{H}- \\ \\ -\text{C}- \\ \end{array}$ | Bonded to either -OH or -O-. |
| 115.1 D | $\begin{array}{c} -\text{C}=\text{O} \quad \text{CH}_3 \\ \quad / \\ \text{C}^*\text{H}=\text{C} \\ \backslash \\ \text{CH}_2- \end{array}$ | |
| 38.1 S | $\begin{array}{c} \\ -\text{C}-\text{CH}_3 \\ \end{array}$ | 1-2 CH ₃ ^s 0-3 CH ₂ ^s 0-2 CH ₃ ^s |
| 39.3 S | $\begin{array}{c} \\ -\text{C}-\text{CH}_3 \\ \end{array}$ | 1-2 CH ₃ ^s 0-1 CH ₂ ^s 1-2 CH ₃ ^s |
| 147.5 S | $\begin{array}{c} \text{CH}_2-\text{CH}_2- \\ / \\ \text{CH}_2=\text{C}^* \\ \backslash \\ \text{CH}-\text{CH}_2- \\ \\ -\text{C}- \\ \end{array}$ | |
| 163.4 S | $\begin{array}{c} \text{CH}_3 \quad \text{O} \\ \backslash \quad / \\ \text{C}^*=\text{CH}-\text{C}-\text{OH} \\ / \\ \text{CH}_2 \\ \\ \text{CH}_2- \end{array}$ | |
| 171.0 S | >C=O |) |
| 171.8 S | >C=O |) |
| | |) Bonds to -OH, -O-, -CH ₃ and -CH= to be distributed between these carbonyls. |

Figure 6. Substructural information derived through analysis of ¹³C data in Table IV and passed to the GENOA program.

Table V. ¹³C Spectrum of a Diterpene Isolated from *Roylea calycina* (Roxb) Briq⁴⁵

| shift, δ | mult ^a | shift, δ | mult ^a | shift, δ | mult ^a | shift, δ | mult ^a |
|----------|-------------------|----------|-------------------|----------|-------------------|----------|-------------------|
| 8.3 | q | 21.6 | t | 51.0 | d | 37.0 | s |
| 16.1 | q | 22.8 | t | 77.3 | d | 43.0 | s |
| 21.1 | q | 25.7 | t | 110.7 | d | 81.5 | s |
| 21.5 | q | 34.9 | t | 138.5 | d | 124.8 | s |
| 27.6 | q | 38.5 | t | 143.0 | d | 170.5 | s |
| | | 41.0 | d | | | 211.2 | s |

^a Multiplicity.

tostrobooides led to several inconsistencies due to the lack of appropriate substructural prototypes in the data base. Interpretation requiring only one-bond substructural models served merely to confirm the presence of a number of substructural features already identified by IR or ¹H NMR spectroscopies. This, then, is an example where the prediction method must serve to reduce the number of structural possibilities.

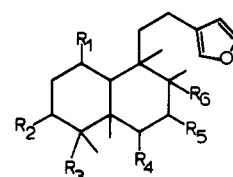
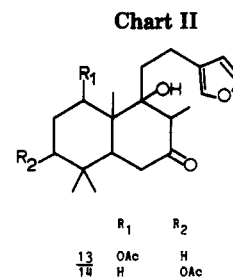
Candidate structures were generated for this unknown by means by the GENOA program.⁴ Although the authors assumed the presence of a labdane skeleton,⁴⁵ we again made the more general assumption that the unknown incorporated a standard diterpene skeleton from the set mentioned previously. Proton resonance and IR data had established the presence of a β-substituted furan ring and an acetoxy group; application of these two constraints eliminated all but the labdane and clerodane skeleton systems. Other proton resonance and IR data established the presence of CHCH₃, three tertiary methyl groups, a tertiary hydroxy group, a ketone (not conjugated with the furan) and the substructure CH₂CH(OAc). The methine in CH(OAc) gave a triplet at 4.75 ppm in the ¹H NMR spectrum suggesting that the neighboring atoms were CH₂ and a quaternary carbon; however, since it is possible that conformational factors could be responsible for the absence of another coupling to the proton, the substructural constraint used allowed for both a quaternary and a methine as the second neighboring atom.

Application of these constraints leads to a final total of 112 structures including both the labdane and clerodane skeletal types. The ¹³C spectra were predicted for these constitutional isomers, most of which incorporated features novel to the data base. Thus, the average shell level used for prediction was only 1.7. The predicted spectra were matched against the observed data to derive their dissimilarity scores (eq 1, 2²⁷) which were then used for structure ranking. These dissimilarity scores ranged from about 25 to 225. Thirty-four structures were selected for further analysis; the structures selected were those with dissimilarity scores below 60 (at which point there was in this example a quite large break in the observed distribution of scores). Twenty-seven of these structures were clerodanes and seven were labdanes. The preponderance of clerodanes among the survivors is partly a reflection of the fact that the data base contained far fewer standard clerodanes than labdanes, and, consequently, predictions for clerodanes had to be based on poorer models which were scored more generously by the ranking functions. At this point the number of candidates was sufficiently small to make it practical to draw them and inspect them for inappropriate substructural features.

Several of the surviving 34 candidates incorporated features incompatible with the proton resonance data. Thus, the structure ranked highest on the predicted ¹³C data (12) was eliminated because it should show a two-proton singlet or an AX system corresponding to the

Table VI. Ranking Results with ^{13}C Spectrum Dissimilarity Scores for the Final 13 Candidate Structures for the Diterpene Isolated from *Roylea calycina* (Roxb) Briq⁴⁵

| struct | dissimilarity score | av shell level |
|--------|---------------------|----------------|
| 24 | 28.3 | 1.9 |
| 16 | 36.7 | 1.9 |
| 17 | 40.5 | 1.8 |
| 22 | 41.1 | 1.8 |
| 25 | 44.6 | 1.9 |
| 13 | 46.7 | 2.3 |
| 19 | 47.6 | 1.8 |
| 21 | 49.7 | 1.8 |
| 14 | 50.1 | 2.6 |
| 23 | 50.2 | 1.8 |
| 15 | 50.5 | 1.8 |
| 18 | 53.1 | 1.8 |
| 20 | 54.5 | 1.8 |



| | R ₁ | R ₂ | R ₃ | R ₄ | R ₅ | R ₆ |
|----|----------------|----------------|----------------|----------------|----------------|----------------|
| 15 | OAc | H | H | =O | H | OH |
| 16 | H | OAc | H | =O | H | OH |
| 17 | OAc | H | OH | =O | H | H |
| 18 | OAc | H | H | H | =O | OH |
| 19 | H | OAc | H | H | =O | OH |
| 20 | H | OAc | OH | H | =O | H |
| 21 | OAc | H | OH | H | =O | H |
| 22 | =O | H | H | OAc | H | OH |
| 23 | H | =O | H | H | OAc | OH |
| 24 | H | =O | H | OAc | H | OH |
| 25 | H | =O | OH | H | OAc | H |

protons on C-12 at around 3.7 ppm in the ^1H NMR spectrum. Such checks against the ^1H NMR data eliminated all but 13 of the remaining candidates. The surviving candidates are structures 13–25 (Chart II). The ^{13}C spectrum dissimilarity scores are summarized in Table VI. Of these structures, two (13 and 14) are labdanes, and the rest are clerodanes. Examples of predicted ^{13}C spectra are given in Table VII for some of these structures.

Stereoisomers were generated, with the assumption of standard configurations for the diterpene skeletons, and the spectrum prediction and ranking procedures were repeated. By use of the stereochemical representations, more precise spectrum predictions can be achieved. A resonance-range prediction for an atom in a given topological (constitutional) environment generally combines data from different configurational forms and thus is broadened. Because for these compounds the prototype substructures used for prediction averaged less than two bond environments, stereochemical factors were rarely represented. Thus, each pair of epimeric clerodane-type structures obtained identical scores; however, the use of stereochemistry did improve the predictions for the two pairs of epimeric labdanes. The improvement in prediction mainly resulted from the fact that, by including stereochemistry, a dis-

inction could be made between the two diastereotopic methyls. If just a constitutional model is used, these two methyls appear the same and are both predicted to resonate at about 22 ppm. They are, however, distinguished in the stereochemical codes with one resonance expected around 27 ppm and the other at about 17 ppm on the basis of methyl groups in the data base in similar environments. The difference in scores was not sufficient to identify the correct stereoisomer with certainty.

In ref 45, only labdane-type structures were considered; the clerodanes could possibly have been excluded on the basis of knowledge about the biosynthetic processes in the plant (*Roylea calycina* (Roxb) Briq) from which the compounds were derived. If such an assumption is made, the spectral data alone suffice to reduce the number of possible structures to just 13 and 14. The clerodane-type structures

Table VII. Observed Spectrum and Predicted ^{13}C Spectra (in δ) for Structures 13, 14, 16, 22, and 24

| obsd | predicted ^a | | | | |
|-----------|------------------------|-----------|-----------|-----------|-----------|
| | 13 | 14 | 16 | 22 | 24 |
| 211.2 (s) | 211.3 (1) | 211.3 (1) | 213.0 (1) | 211.3 (1) | 211.3 (1) |
| 170.5 (s) | 170.6 (3) | 170.4 (4) | 170.2 (3) | 170.6 (3) | 170.6 (3) |
| 124.8 (s) | 125.5 (4) | 125.5 (4) | 125.5 (3) | 125.5 (3) | 125.5 (3) |
| 81.3 (s) | 77.0 (1) | 80.3 (2) | 80.2 (1) | 80.2 (1) | 80.2 (1) |
| 43.0 (s) | 46.5 (0) | 41.2 (1) | 51.2 (1) | 41.2 (1) | 41.2 (1) |
| 37.0 (s) | 33.2 (2) | 37.6 (2) | 41.2 (1) | 40.7 (1) | 40.7 (1) |
| 143.0 (d) | 143.1 (4) | 143.1 (4) | 143.1 (4) | 143.1 (4) | 143.1 (4) |
| 138.5 (d) | 138.7 (3) | 138.7 (3) | 138.7 (3) | 138.7 (3) | 138.7 (3) |
| 110.7 (d) | 110.9 (4) | 110.9 (4) | 110.9 (3) | 110.9 (3) | 110.9 (3) |
| 77.3 (d) | 79.5 (1) | 80.5 (3) | 74.7 (1) | 79.5 (1) | 79.5 (1) |
| 51.0 (d) | 53.1 (1) | 53.1 (1) | 53.1 (1) | 61.3 (1) | 48.7 (1) |
| 41.0 (d) | 48.7 (1) | 48.7 (1) | 38.3 (1) | 36.3 (1) | 44.5 (2) |
| 38.5 (t) | 39.2 (3) | 32.8 (2) | 48.0 (1) | 39.7 (1) | 39.7 (1) |
| 34.9 (t) | 30.8 (2) | 31.7 (3) | 36.8 (1) | 36.8 (1) | 36.8 (1) |
| 25.7 (t) | 30.7 (2) | 30.8 (2) | 26.7 (1) | 33.9 (1) | 33.9 (1) |
| 22.8 (t) | 25.6 (2) | 25.6 (2) | 20.7 (2) | 26.7 (1) | 22.6 (2) |
| 21.6 (t) | 20.7 (3) | 20.1 (4) | 20.1 (2) | 20.7 (1) | 20.7 (2) |
| 27.6 (q) | 27.5 (3) | 22.0 (3) | 28.4 (2) | 28.4 (2) | 28.4 (2) |
| 21.5 (q) | 27.5 (3) | 22.0 (3) | 21.4 (4) | 21.2 (4) | 21.2 (4) |
| 21.1 (q) | 21.2 (4) | 21.2 (4) | 16.5 (2) | 16.4 (2) | 16.4 (2) |
| 16.1 (q) | 19.6 (1) | 16.4 (2) | 16.4 (2) | 16.3 (2) | 12.8 (2) |
| 8.2 (q) | 12.2 (2) | 12.2 (2) | 12.3 (2) | 12.8 (2) | 12.2 (2) |

^a The numbers given in parentheses after each shift value define the shell level of the prototype substructure used to make that prediction.

are definitely eliminated by the results of the first chemical step used in the original characterization of this structure.⁴⁵ Dehydration of the compound led to an α,β -unsaturated ketone with the original secondary methyl appearing as the sole vinyl methyl group in the product. The remaining candidate clerodane-type structures are incompatible with the structural constraints implied by this experiment. On the basis of the spectral and chemical transform data, the unknown was eventually characterized⁴⁵ as being structure 14, with the configuration at C-3 being uncertain.

Our assignment of resonances based on the prototype substructures in the current data base differs somewhat from the tentative assignment reported in ref 45. The use of the data base and special functions for assisting in assigning resonances are described elsewhere.^{27,46}

Discussion

The two examples chosen to illustrate the use of these programs were determined by the availability of extensive, reliable ¹³C data for certain classes of diterpenes and the fact that new compounds from these classes continue to be identified in various terrestrial plants. These factors helped to make it possible to undertake analyses with what is still a quite restricted data base. More elaborate analyses can be attempted as the data base is expanded.

With the current data base, it is common for new structures to contain novel substructural features as yet unrepresented in the files. The atoms in such structures will be matched by at best one-bond prototype environments, and, of course, occasionally *no* relevant prototype will exist. For very small molecules, C₁₀ etc., it is sometimes possible to "solve" the structure by using the interpretive procedures and requiring only one-bond atom environments. For most compounds of biological interest, attempted interpretation by using a one-bond requirement will not yield enough constraints, and structure generation will be impractical. Attempts at spectral interpretation requiring more precise matching levels than are in fact represented by the prototype substructures in the data base leads, in general, to inconsistencies. In such instances the substructures retrieved for some resonance will imply bonding constraints on the corresponding atom in the structure that are incompatible with the bonding constraints derived for other atoms. Such inconsistent interpretations result in insufficient bonds available to satisfy some atom's valence and, consequently, are readily detected by the program and reported to the investigator. It may be possible to find consistent interpretations if extended shell matches are required. In this instance it is possible to generate a set of structural candidates which does not contain the correct structure. Although we have not encountered this circumstance, it is always wisest to be conservative in the requirements for shell matches and to supplement the ¹³C interpretations with structural information available from other sources when actually generating structures.

The procedures for spectrum prediction and ranking are less sensitive to limitations of the data base. The lack of good prototype substructures in the data base merely limits the extent to which inappropriate candidate structures can be identified and eliminated from further consideration. However, even our limited data base is sufficient to provide prototype substructural codes out to shell levels sufficient to eliminate many structural candidates which would remain if *only* data from other spectroscopic techniques (¹H

NMR, IR) were used to provide constraints on the potential structural variety.

The main limitation of the spectrum prediction and structure ranking scheme is of course the prerequisite of a finite set of candidate structures. The procedures for spectrum prediction and ranking of hypothesized structures seem generally satisfactory, though more sophisticated spectrum-matching functions are desirable. As the reference files are expanded, more substructural environments will be represented, and predictions for different hypothesized candidates will be based more often on models of similar specificity. This will tend to enhance the discriminatory power of the spectrum-matching and ranking procedures, because it will become less common to have to retain candidates simply because the data base did not provide adequate codes for a precise spectrum prediction.

It is not essential to have a complete ¹³C spectrum with line multiplicities for application of the spectrum prediction and ranking procedures. Modification of the current program would allow the use of partial spectral data. However, the discriminatory power of these procedures would in general be significantly reduced.

The interpretive approach will invariably depend on having many reference compounds in the file that are closely related to an unknown. Though the success of the interpretation approach is limited by the quality of the data base, it does allow more extensive use of ¹³C data than either conventional classification of carbons by hybridization and hydrogen substitution or file-search spectrum recognition. When demanding matching criteria are set for shifts and substructures, it is possible to use the interpretation functions as a form of spectrum-recognition procedure; however, conventional file-search methods are considerably more efficient at such recognition tasks.

In some respects, the current interpretive approach to processing ¹³C data is overspecialized and overautomated. Though it is of interest to find that ¹³C data alone can suffice to identify a moderately complex structure, it would be an unusual analytical problem in which only ¹³C data were available. Substructures inferred from ¹H NMR, IR, or chemical data would provide valuable additional constraints on the range of substructural codes that should be retrieved. Certain limited use of such substructural inferences is already possible. Though the actual mechanisms are somewhat clumsy, it is possible to incorporate any substructural constraint involving carbons that can be associated with specific resonances. Further development will allow for some more general substructural constraints to be used. There are, however, some intrinsic limits, and the full generality of the substructural constraints in GENOA cannot be attained. Other proposed developments will eventually allow consistent treatment of stereochemical information throughout the spectrum-interpretation and structure-generation processes.

Problems consequent upon the interpretation functions attempting to solve the structure and assign the spectrum simultaneously have been noted earlier. Such problems are really just one instance of where the current programs are simple not sufficiently interactive. Some mechanisms exist for the user to inspect intermediate results and apply additional constraints, but these mechanisms currently lack flexibility. Further development of these programs is intended to lead to a more fully interactive approach in which spectrum interpretation is achieved jointly by cooperation of the investigator and the programs.

Successful application of the programs to a given structural problem outside the scope of the current data

(46) C. W. Crandell, A. Lavanchy and N. A. B. Gray, unpublished results.

base, which consists largely of oxygen-containing natural products, will, in general, require the creation of new, class-specific data bases. This has already been accomplished for a series of diterpenoid alkaloids.³⁸ We are willing to collaborate with other interested investigators in such efforts (See Experimental Section).

Experimental Section

These programs are implemented in the ALGOL-like BCPL program language⁴⁷ on a Digital Equipment Corp. KI-10 computer at the SUMEX-ALM computer facility at Stanford. The programs are available to an outside community of investigators via a nationwide computer network to the limit of available resources. Export of the programs to other DEC PDP-10 or PDP-20 systems

(47) M. Richards and C. Whitby-Stevens, "BCPL—The Language and Its Compiler", Cambridge University Press, Cambridge, 1979.

or other computers supporting BCPL (e.g., IBM-370) is possible. However, additional work remains before the programs become polished enough for mass export. Meanwhile, within the limits of our resources, we are prepared to collaborate in the ¹³C-based solution of nontrivial structure problems for outside investigators who lack appropriate computer facilities.

Acknowledgment. We thank the National Institutes of Health (Grant No. RR-00612 and AM-04247) for their generous financial support and the United Kingdom Science Research Council for a grant to N.A.B.G. (B/RF/4955). Computer resources were provided by the SUMEX facility at Stanford University under National Institutes of Health Grant No. RR-0785.

Registry No. 4, 73809-96-8; 5, 75991-61-6; 11, 63399-37-1; 13, 75919-20-9; 14, 75947-44-3; 16, 75919-21-0; 22, 75919-22-1; 24, 75919-23-2.

Carbanions. 5.¹ Preparation and ¹H and ¹³C NMR Spectroscopic Structural Study of the 4-Hydridopyridyl Anion and 4,4'-Bis(hydridopyridyl) Dianion. Absence of Homoazacyclopentadienyl Ion Character

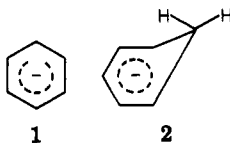
George A. Olah* and Ronald J. Hunadi

Hydrocarbon Research Institute and Department of Chemistry, University of Southern California, Los Angeles, California 90007

Received November 12, 1980

The 4-hydridopyridyl anion (3) was prepared and studied by NMR spectroscopy. By analogy with the cyclohexadienyl anions, ion 3 was shown to be planar with no 1,5-homoaromatic overlap occurring. Temperature-dependence studies showed that there is no change in the structure of 3 down to -40 °C and consequently no puckered form could be frozen out. At room temperature anion 3 was slowly converted into the 4,4'-bis(hydridopyridyl) dianion (7). The structure of 7 was also confirmed by its independent preparation.

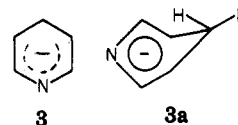
Our preceding study of the cyclohexadienyl anion 1 by NMR spectroscopy as well as by MINDO/3 calculations concluded that ion 1 was a planar nonhomoaromatic species with no significant contribution, if any, by the homocyclopentadienyl anion 2.² The methylene protons



were equivalent even at -60 °C (at 300 MHz) and thus no indication for 2 was obtained. Recently Haddon³ calculated that structure 1 was the major energy minimum for the C₆H₇ potential-energy surface and was 36 kcal/mol lower in energy than 2.

Bodor and Pearlman⁴ in 1978 reported the results of their MINDO/3 study of dihydropyridine anions and related dihydropyridyl species. They concluded that the

4-hydridopyridyl anion (3) should be planar (within 0.5°)



with the charge delocalized over the five atoms. Although these calculations on bond angles and geometry were similar to those for the 1,4-cyclohexadienyl anion, no experimental data were available. The possibility of 1,5-homoaromatic overlap in the case of the 4-hydridopyridyl anion (3) was difficult to completely rule out, at least as possible contribution of the homoazacyclopentadienyl anion (3a) to the structure of 3.

Recently Fraenkel et al.⁵ reported their work on the generation of spirodihydroaromatic anions and in this study reported the ¹³C NMR shifts of 4,4-dimethyl-1-lithio-1,4-dihydropyridine (4).⁴ They concluded, from the equivalence of the methyl protons in the ¹H NMR spectrum and the methyl carbons in the ¹³C NMR spectrum

(1) For Part 4 see: Olah, G. A.; Hunadi, R. J. *J. Am. Chem. Soc.* 1980, 102, 6989.

(2) Olah, G. A.; Asensio, G.; Mayo, H.; Schleyer, P. v. R. *J. Am. Chem. Soc.* 1978, 100, 4347.

(3) Haddon, R. C. *J. Org. Chem.* 1979, 44, 3608.

(4) Bodor, N.; Pearlman, R. *J. Am. Chem. Soc.* 1978, 100, 4946.

(5) Rizvi, S. Q. A.; Foos, J.; Steel, F.; Fraenkel, G. *J. Am. Chem. Soc.* 1979, 101, 4488.

(6) Birch, A. J.; Karakhanov, E. A. *J. Chem. Soc., Chem. Commun.* 1975, 480.

(7) Foos, J.; Steel, F.; Rizvi, S. Q. A.; Fraenkel, G. *J. Org. Chem.* 1979, 44, 2522.