enumeration of particular structures. Such extensions to the relaxation procedure are important subjects for future research.

ACKNOWLEDGMENT

The support of the National Science Foundation (Grant MCS-79-23422) is gratefully acknowledged, as is the help of Janet Salzman in preparing this support.

REFERENCES AND NOTES

(1) E. V. Krishnamurthy and M. F. Lynch, "Analysis and Coding of Generic Chemical Formulae in Chemical Patents", *J. Inf. Sci.*, in press.

(2) M. F. Lynch, J. M. Barnard, and S. M. Welford, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy", *J. Chem. Inf. Comput. Sci.*, **21**, 148 (1981).

(3) J. M. Barnard, M. F. Lynch, and S. M. Welford, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL: A Formal Language for the Description of Generic Chemical Structures" *J. Chem. Inf. Comput. Sci.*, **21**, 161 (1981).

(4) S. M. Welford, M. F. Lynch, and J. M. Barnard, "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **21**, 161 (1981).

(5) L. Kitchen and A. Rosenfeld, "Discrete Relaxation for Matching Relational Structures", *IEEE Trans. Syst. Manage. Cybern.*, **SMC-9**, 869-874 (1979).

(6) H. G. Barrow and J. M. Tenenbaum, "MSYS: A System for Reasoning about Scenes", SRI AI Center, Menlo Park, CA, 1976, Tech. Note 121.

(7) J. R. Ullman, "An Algorithm for Subgraph Isomorphism", *J. Assoc. Comput. Mach.*, **23**, 31-42 (1976).

(8) D. Waltz, "Understanding Line Drawings of Scenes with Shadows" in "The Psychology of Computer Vision", P. H. Winston, Ed., McGraw-Hill, New York, 1975, 19-92.

(9) R. M. Haralick and L. G. Shapiro, "The Consistent Labelling Problem: Part I", *IEEE Trans. Pat. Anal. Mach. Intel.*, **PAMI-1**, 173-184 (1979).

(10) R. M. Haralick and L. G. Shapiro, "The Consistent Labelling Problem: Part II", *IEEE Trans. Pat. Anal. Mach. Intel.*, **PAMI-2**, 193-203 (1980).

(11) A. K. Mackworth, "Consistency in Networks of Relations", *Artif. Intel.*, **8**, 99-118 (1977).

(12) J. Gaschnig, "Experimental Case Studies: Backtrack vs Waltz-Type vs New Algorithms for Satisficing Assignment Problems", in Proceedings of the Second National Conference of the Canadian Society for Computational Studies of Intelligence, University of Toronto, Toronto, Ontario, July 1978, pp 268-277.

(13) J. J. McGregor, "Relational Consistency Algorithms and Their Application in Finding Subgraph and Graph Isomorphisms", *Inf. Sci.*, **19**, 229-250 (1979).

(14) C. Rieger, J. Bane, and R. Trigg, "ZMOB: A Highly Parallel Multiprocessor", University of Maryland, College Park, MD, May 1980, Computer Science Tech. Report 911.

(15) C. Rieger, R. Trigg, and J. Bane, "ZMOB: A New Computing Engine for AI", University of Maryland, College Park, MD, March 1981, Computer Science Tech. Report 1028.

(16) C. Rieger, "ZMOB: Hardware from a User's Viewpoint", University of Maryland, College Park, MD, April 1981, Computer Science Tech. Report 1042.

Structure Evaluation Using Predicted ^{13}C Spectra[1]

CHRISTOPHER W. CRANDELL, NEIL A. B. GRAY, and DENNIS H. SMITH*

Department of Chemistry, Stanford University, Stanford, California 94305

Received August 14, 1981

A computer program is described for predicting ^{13}C NMR spectra of organic compounds and for determining the similarity of the predicted spectra to an observed spectrum. The program utilizes a data base containing representations of the stereochemical substructural environments of resonating nuclei, together with their chemical shifts. Given the observed spectrum of an unknown compound and a set of structural candidates for the unknown, the predicted spectra are matched with the observed spectrum, and a score reflecting the degree of matching is calculated. Alternative methods for matching spectra and computing scores are discussed and evaluated using several examples. A matching and scoring function which takes into account the limitations of the data base has proven to yield the best performance.

INTRODUCTION

In recent years, a number of computer programs have been developed that can construct hypothetical candidate structures for an unknown molecule.[2-6] Gribov[7] and Hippe[8] have recently reviewed such programs and related computer systems. Typically, these programs work by taking a set of substructural fragments, either identified by the chemist or automatically inferred from spectral data and assembling these in all possible ways. The chemist can thus be provided with a set of candidate structures each compatible with all of the more readily interpretable chemical and spectral data. Most structure generation programs work solely in terms of constitutional (topological) isomers; recently, however, algorithms for the constrained generation of configurational stereoisomers have been perfected, thus allowing for more complete structure elucidation.[9]

Such a set of candidate structures, as constitutional or stereoisomers, can in itself be of value to the chemist. Examination of the possible structures can help identify additional spectral or chemical experiments that would resolve remaining alternatives. However, it is also possible for the computer systems to assist in the process of evaluation of the candidates. Typically, programs for candidate structure evaluation are concerned with predicting *spectral* properties for candidate structures and comparing predicted and observed spectra. Differences between predicted and observed spectral properties can be used directly to eliminate candidate structures, or more conservatively, such differences can be captured in some measure of spectral (dis)similarity that is then used for rank ordering the candidates.

Spectral prediction and evaluation algorithms must be capable of processing large numbers of structures of closely related form. This necessarily constrains the type of approach that can reasonably be adopted. In principle, it is possible to use ab initio or semiempirical quantum mechanical methods to compute certain spectral properties of candidate structures, and indeed some limited experiments have been made using MINDO-level semiempirical quantum mechanical methods to calculate geometries for small candidate molecules produced by the STREC system.[10] However, quantum mechanical approaches have not played a major role in evaluation of candidate structures due both to their high computational cost and generally poor ability to discriminate between closely related isomers. Instead, most approaches to spectral prediction are based upon simple empirical models and correlation techniques.

Methods for spectrum prediction–structure evaluation using empirical correlations all have intrinsic limitations and biases. *These limitations must be well characterized before such empirical methods can appropriately be applied to problems of structure elucidation.*

Recently, we reported a method for ^{13}C spectrum prediction that relies upon the use of a data base of substructure/shift combinations.[11,12] This data base is employed as a kind of very detailed correlation chart giving the expected shift range for carbon atoms in particular stereochemical environments. The spectra of hypothesized candidate structures predicted on the basis of information in the data base are "fuzzy". Each carbon atom of the structure is ascribed not a specific chemical shift but rather a spectral range within which its resonance should be observed. For some candidate structures, the data base may contain quite detailed substructural models and the spectral predictions may be quite precise; for other candidate structures the data base may lack appropriate models. This results in broadened predicted shift ranges for atoms and thus a poorly defined spectrum.

When evaluating different hypothesized structures, it is necessary to match these predicted fuzzy spectra with the observed line spectrum and to compute some score expressing the quality of the spectral match. For this spectrum-prediction/structure-evaluation approach to be of value in routine applications, it is important that the computed matching score not be overly sensitive to small variations in the quality of predictions and that the method suitably take into account any intrinsic bias toward reference structures used as sources for shift information in the data base.

This paper reports results of experiments undertaken to determine appropriate methods for matching spectra and scoring the quality of the spectral match. These experiments use limited, clearly defined data sets and have been designed to allow comparison with earlier work.[13]

EXISTING SPECTRAL-BASED METHODS FOR STRUCTURE EVALUATION

Mass spectral data have been used for several years as the basis for programs for structure evaluation. One factor making mass spectrometry a useful technique is the relative insensitivity of mass spectra to the configurational and conformational stereochemistry of molecules and, consequently, its suitability for use with constitutional isomers created by the structure generators, which ignore stereochemistry, summarized above. Recent developments have resulted in two approaches to prediction of mass spectra. Rule-based procedures[14] can be devised that are capable of making fine distinctions among closely related isomers within specified compound classes. A more generally applicable approach to structure evaluation uses the so called "half-order theory" of mass spectrometry.[15] The half-order theory does not incorporate all the factors that influence ion intensity patterns. Consequently, the measure used for structure ranking is an estimate of the ease with which an observed spectrum can be rationalized in terms of simple fragmentation processes operating on a given structure, rather than a conventional spectral matching coefficient.

Prediction and evaluation of IR spectra have also been demonstrated to be potentially useful approaches to structure evaluation. Gribov et al.[16] have developed methods for predicting IR spectra of candidate structures and for computing similarity coefficients for predicted and observed spectra that can be employed for structure ranking.

Munk et al. have utilized ^{13}C data to eliminate inappropriate candidate structures produced by their CASE program.[17] Their approach is based upon analyzing the topological symmetry of a structure and hence determining the expected number of distinct signals in the carbon resonance spectrum. Such an approach has limitations. Most notably, the topological symmetry group is inappropriate for determining magnetic equivalence because topological equivalences are frequently broken by configurational or conformational factors, and indeed, the proposed algorithm has to handle diastereotopic methyl groups as special cases. In addition, chance degeneracy of signals could seriously mislead such an analysis. However, Munk et al. have shown that their simple analysis in some cases provides remarkable discrimination among members of a set of candidate structures.[17]

Other approaches to ^{13}C spectral analysis have involved devising a parameterized function that relates the chemical shift to a number of structural features. Notable examples of such analyses are those by Lindeman and Adams[18] for alkanes and Eggert and Djerassi for amines.[19] Although most of these studies have used simple structural parameters like connectivity and configuration, the use of geometrical structural properties has also been demonstrated.[20] Some of these schemes have been automated and can be used for structural analysis of specific classes of compounds.[21] However, such methods cannot act as the basis for any general scheme for ^{13}C spectral prediction and ranking of arbitrary structures.

Mitchell and Schwenzer have proposed a "production rule" method as an alternative to conventional parametrized linear functions.[13] Production rules define substructural templates and associated shifts; spectral prediction is achieved by matching the available templates to the atoms in a molecule and assigning each atom a shift range defined as the intersection of the ranges given in its matching substructural templates. The emphasis of the Mitchell/Schwenzer work was rather more on the development of the production rules themselves, as a example of "Artificial Intelligence" methods for concept formation and generalization, than on development of a practical tool for structure evaluation.

METHOD

We have previously reported procedures both for ^{13}C spectrum interpretation and for ^{13}C spectrum prediction and structure evaluation.[11,12] These spectrum interpretation and prediction procedures are used in conjunction with the GENOA and STEREO programs for isomer generation.[5,9] These spectral analysis procedures rely on the existence of a data base of substructures and chemical shift data.

This data base represents a compilation of information from standard reference compounds. In the data base the substructural environment of a resonating nucleus is represented out to a four bond radius (inclusion of δ-substituent effects). The correlated chemical shift data gives the mean, minimum, and maximum shifts observed for atoms in such substructural environments among the standard reference compounds. The substructural representation incorporates *both molecular constitution and configurational stereochemistry.*

The present data base contains information on about 21 000 substructures (a resonating carbon and its environment out to maximum of a four-bond radius) gleaned from approximately 1500 compounds (predominantly steroids, terpenoids, and alkaloids). Consistent shift ranges are found for similar substructures in different structures as illustrated by the data in Table I. For example (Table I), the 3576 substructures, which occur more than once in the data base, show a range

Table I. Statistics on Contents of the Data Base[a]

bond radius	no. of substructures	no. of multiple instances	mean range, ppm	standard deviation, ppm
α	1 073	753	8.08	2.90
β	5 817	2 675	2.56	1.10
γ	11 393	3 576	0.98	0.48
δ	15 853	3 494	0.51	0.27

[a] Includes the number of unique substructures, the number which occur more than once, and the mean and standard deviation of the observed chemical shifts.

of shifts of about 1 ppm if they are identical in constitution and stereochemistry out to at least a three-bond radius.

The structure evaluation procedure, as applied to constitutional isomers created by GENOA or stereoisomers created via STEREO, consists of four basic steps: (1) prediction of the ^{13}C spectrum for each candidate, (2) matching each predicted spectrum to the observed spectrum, (3) computing a score for the matching, and (4) rank ordering each candidate based on the score computed.

Spectrum Prediction. The spectrum prediction process employing a data base has been described previously.[11] Briefly, the substructural environment for each carbon nucleus of a given candidate structure is analyzed, the relevant model retrieved from the data base, and the carbon assigned a chemical shift *range* based upon the correlated chemical shift data. Typically, some of the atoms of a candidate structure will be in relatively novel structural environments, and the data base will not include the corresponding four-bond substructural models. Chemical shift predictions for such atoms have to be based on a generalization procedure. Our prediction program derives generalizations as required by trying successively the four-, three-, two-, and one-bond substructural environments for a resonating atom. More general substructure environments are correlated with shift data derived from all the more specific substructures that they subsume. Each such more generalized model is associated with an increasingly wider shift range and a more poorly defined distribution of shifts. The result of the prediction process, as applied to a given candidate structure, is a fuzzy spectrum in which each carbon atom is characterized by a resonance range rather than a specific resonance shift.[11]

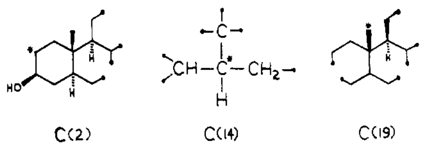
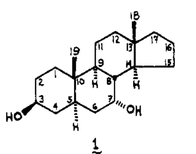
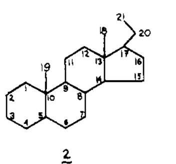


Figure 1. Substructural environments obtained from the data base for C(2), C(14), and C(19) of 5α-androstane-3β,7α-diol (**1**). Note that the atom in question is starred and that the dots represent nonhydrogen attachments.

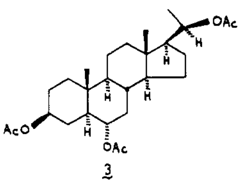
An example of such a "fuzzy" predicted spectrum for 5α-androstane-3β,7α-diol (**1**) is given in Table II. The ranges in predicted shift values vary widely for different atoms. Some atoms, e.g., C(2), are in substructural environments which are well represented in the data base. For such atoms substructural models are found corresponding to shell 3 or 4 substructures (column 3), and relatively narrow shift ranges are predicted (columns 4 and 5). Other atoms, e.g., C(14), have few prototypes in the data base and predictions must be made on the basis of shell 1 environments. Inevitably such predictions lead to broad ranges [33 ppm for C(14)]. In still other cases, such as C(19), predictions may be derived from substructural models in two different configurational forms, and the resulting range may again be large even though a shell 3 model was found in the data base (the shell 4 environment is needed to define the relative configurations of C(5) and C(10) for C(19),[12] but no shell 4 prototype of C(19) in **1** exists in the data base). Figure 1 illustrates the environments of C(2), C(14), and C(19) on which the predictions in Table II are based.



1



2



3

Matching Predicted and Observed Resonances. The next step in the procedure is to match the predicted and observed spectra, i.e., to place the set of resonance lines of the observed spectrum into correspondence with the set of resonance ranges of the predicted spectrum. Limitations in both the prediction

Table II. Predicted Spectrum for 5α-Androstane-3β,7α-diol (1)[a]

			predicted spectrum					obsd spectrum	
atom[b]	mult[c]	shell[d]	res_{min}[e]	res_{max}[f]	mean[g]	SD[h]	no. of res[i]	shift[j]	mult[k]
13	s	2	40.2	41.7	40.8	0.2	23	40.9	s
10	s	2	34.5	38.2	35.7	0.6	104	35.7	s
3	d	3	70.0	72.3	71.1	0.5	37	71.1	d
7	d	3	68.1	68.1	68.1	0.0	1	68.3	d
14	d	1	26.7	59.7	50.1	6.6	1002	48.7	d
9	d	3	46.4	46.4	46.4	0.0	1	46.2	d
5	d	2	41.2	42.1	41.8	0.3	7	40.0	d
8	d	2	39.7	40.0	39.8	0.1	6	37.1	d
17	t	3	38.0	40.6	40.3	0.5	21	40.3	t
12	t	3	38.5	39.3	38.9	0.1	18	38.4	t
4	t	2	35.5	40.6	38.1	1.2	38	37.8	t
1	t	3	36.5	37.5	37.1	0.2	15	36.9	t
6	t	2	36.8	36.8	36.8	0.0	1	36.4	t
2	t	4	31.2	32.5	31.7	0.4	15	31.5	t
15	t	2	24.8	27.7	25.5	0.5	31	24.9	t
11	t	2	18.1	27.3	21.2	1.0	296	21.1	t
16	t	3	20.1	21.2	20.5	0.2	29	20.5	t
18	q	3	17.3	17.8	17.5	0.1	23	17.3	q
19	q	3	11.2	24.3	16.7	5.7	104	11.3	q

[a] Sorted by multiplicity groups and arranged in decreasing magnitude of resonance values. All resonance values in ppm with respect to Me_4Si. [b] Atom number of resonating carbon—see 1. [c] Predicted multiplicity. [d] Shell level at which prediction made; the greater the shell level the more of the substructural environment of the atom is specified. [e-i] The predicted minimum, maximum, mean, standard deviation, and number of the resonances in the data base for the substructural environment of the atom at this shell level. [j] Observed resonance. [k] Observed multiplicity.

procedure and the observed data have to be considered in this process.

The matching process is essentially equivalent to one of using chemical shift analogies to assign observed resonances to the constituent atoms of the candidate structure. While a few resonance assignments may be unambiguous, most atoms in a given candidate structure will have predicted ranges encompassing several observed resonances. For example, in the data given in Table II for **1**, the resonance range associated with C(6) includes all the observed triplet resonances. Consequently, any of the observed triplets could be assigned to C(6). As a consequence of better substructural models in the data base (higher shell values, column 3), other CH_2 resonances are associated with narrower predicted ranges [e.g., C(12), C(17)]. It would, in general, be appropriate for these resonances to be matched to the observed triplets first; subsequently, resonances with wider predicted ranges could be assigned. Most of the matching functions investigated make use of the precision of the predictions (as reflected either through prediction ranges or through shell levels) and accord greater importance to matching resonances for which predictions were derived from detailed substructural models.

Another factor which must be taken into account when matching spectra is that an observed spectrum may be incomplete in some manner. For example:

(A) *Missing Resonance Lines.* Lines may be missing due to long relaxation times for some quaternary carbons or unrecognized accidental degeneracy of two or more resonances.

(B) *Missing Multiplicities.* Multiplicities may be unavailable in the absence of an SFORD spectrum (lack of instrument time or inadequate sample).

(C) *Incomplete Multiplicity Data.* The complexity of the SFORD spectrum, or use of a special pulse sequence,[22,23] may make it possible to determine only the parity (even/odd) of the multiplicity of one or more resonances.

The observed data, and also the predicted resonances, are grouped on the basis of available multiplicity data. The matching procedures are most efficient when complete multiplicities are available for the observed spectrum because then the matching functions will only analyze combinations of observed and predicted resonance of the same multiplicity. When less complete observed data, the program will attempt assignments within the groups of even and odd multiplicity or, if necessary, the entire observed spectrum is considered as one group to be matched to the predicted spectrum.

The next step in the analysis is to match individual observed resonance lines with specific predicted resonance ranges in each multiplicity group. A number of methods have been considered.

(A) *MAG*: Simple intercomparison of observed and mean predicted shifts. MAG orders the means (column 6, Table II) of the predicted resonance ranges on the basis of their magnitudes. The predicted and observed resonances are then set into correspondence simply on the basis of their orders within their respective multiplicity groups. This represents the least sophistcated method of matching a predicted to an observed spectrum, but it is also generally the first step in an assignment process. This matching corresponds to that presented for **1** (Table II). If observed resonances are missing for any reason (see above), then a "best" matching is determined by using the algorithm summarized below.

(B) *RANGE*: Attempts to guarantee that every observed shift is assigned to an enclosing predicted range. RANGE is based upon a suggestion[24] in which predicted resonance minima and maxima are used to order predicted spectra. The method involves starting from low field (large shift values) and ordering on the basis of the minimum predicted resonances, moving from low to high field. If two minima are equal, the one with the larger maximum is chosen. The objective is to attempt to arrange that every observed resonance lies in one of the predicted ranges. If observed resonances are missing, this matching function cannot be used since it is not straightforwardly adaptable to the best assignment algorithm described below.

(C) *SMBEL*: Schwenzer/Mitchell range-weighted similarity assignment. SMBEL is based on a slightly modified version of the scoring function used by Schwenzer and Mitchell.[13] A "best" matching is found by maximizing the value of eq 1, using the algorithm summarized below

$$\text{SMBEL} = \sum_{i=1}^{\text{numlines}} [(1 - \text{err}_i/\text{range}_i)/\text{range}_i] \qquad (1)$$

where numlines is the number of observed resonances, err_i = the absolute value of the difference between the ith observed resonance and the ith matched predicted resonance (mean of resonance range), range_i = the absolute difference between the predicted minimum and maximum for the ith carbon in question with a default value (2.0) used if the range is too small. Note that narrow predicted ranges carry more weight than wide predicted ranges.

(D) *SDIS2*: shell weighted dissimilarity assignment. SDIS2 is an empirically devised function reflecting our perceptions of the factors in a matching scheme that are necessary to accommodate the limitations of both the observed spectra and the data base. The ordering is found by minimizing the "disbelief" given by eq 2, using the algorithm summarized below

$$\text{SDIS2} = \sum_{i=1}^{\text{numlines}} [\text{shell}_i \times \text{err}_i^2/8.0] \qquad (2)$$

where numlines is the number of observed resonances, err_i is as defined above for SMBEL, shell_i is the shell level at which the ith prediction was made (i.e., the bond-radius value defining the quality of the match between an atom's substructural environment in a candidate structure and the atom environment as defined in the substructural model used as the basis for prediction). Eight (8.0) is a simple scale factor to reduce the magnitude of the resulting score. This function puts less emphasis on predicted resonances based on poor models (shell = 0 or 1) and, at the same time, increases the disbelief if the values of an observed and a matched predicted mean resonance are quite different (large err). The basic idea is to simultaneously minimize the differences between predicted resonances and matched observed resonances, all the while paying the most attention to the predictions based on good models (shell = 3 or 4).

The SMBEL and SDIS2 matching functions depend upon predicted ranges or the shell level at which predictions are made. Thus, even if all resonances are observed, there may be more than one way to assign predicted to observed resonances, each of which yields a different score. The same situation pertains to these functions and the MAG function when resonances are missing in the observed spectrum. An algorithm has been designed to explore all possible matchings in these circumstances to determine the "best" assignment (lowest value of MAG and SDIS2, highest value of SMBEL).

Algorithm. Within each grouping of predicted and observed resonances (e.g., all quartets or all even-parity resonances), a matrix of scores is constructed; the ith, jth entry in this matrix is the contribution to the (dis)similarity score that would result if predicted resonance i were to be matched to observed resonance j. The contribution to the score is taken as the absolute value of the difference in resonance shift for MAG matching and as defined in eq 1 and 2 for SMBEL and SDIS2, respectively. This matrix is square if all observed

Table III. Data for an Artificial Observed Spectrum and a Corresponding Predicted Spectrum[a]

obsd	predicted spectrum			
spectrum	mean	res_{min}	res_{max}	shell
50.0	55.0	50.0	60.0	1
40.0	50.0	49.0	51.0	4
30.0	22.0	13.0	29.0	2
20.0	15.0	8.0	29.0	1
10.0	10.0	9.0	11.0	4

[a] The Data are contrived to indicate the differences between the functions for spectrum matching. For an explanation of terms, see the footnote to Table II.

Table IV. Matching of Predicted to Observed Resonances by the Four Matching Functions

obsd	predicted resonances			
spectrum	MAG	RANGE	SMBEL	SDIS2
50.0	55.0	55.0	50.0	50.0
40.0	50.0	50.0	55.0	55.0
30.0	20.0	20.0	15.0	20.0
20.0	15.0	10.0	20.0	15.0
10.0	10.0	15.0	10.0	10.0

resonances are present in the observed spectrum, otherwise rectangular.

A depth first search is used to explore all paths through the matrix (i.e., all combination of assignments) and thus find the "best" assignment as the one yielding the least dissimilarity or greatest similarity score. In the worst case (the spectrum of an n carbon atom structure with no multiplicity information) this will involve finding the best path through an $n \times n$ matrix. The scope of the search is reduced by treating observed resonances with similar shifts (differing by less than 0.3 ppm) as essentially equivalent and not considering the effect of interchanging their assignments. Once a possible assignment has been derived, its overall score, and also the partial scores corresponding to incomplete assignments, can be used to limit further exploration. Exploration of a subsequent possible partial assignment may be terminated if the matching score along the current search path exceeds the best score obtained in previous search paths. Even so, as illustrated in subsequent examples, the absence of multiplicity information can result in a very costly search.

There are significant differences in the matching functions as applied to various distributions of predicted and observed spectral data. These differences are best illustrated through the contrived data given in Tables III and IV. The artificial observed and predicted "spectra" (lacking multiplicities) are given in Table III; the results from the four matching processes are in Table IV. The predicted resonances were chosen specifically to demonstrate the variations in assignments obtained with the functions (see columns 2–5 of Table IV). Thus, depending on the matching function, the following pairs of observed resonances can be interchanged among the predicted resonances: 55.0 and 50.0, 20.0 and 15.0, and 15.0 and 10.0.

The MAG function ignores the quality of the predicted spectra, simply ordering and matching resonances according to their magnitudes (column 2, Table IV).

The RANGE function interchanges the matchings of the observed 10- and 20-ppm (column 3, Table IV) resonances because the predicted minimum resonance for the 15-ppm mean reasonance is lower than that for the 10-ppm mean resonance (columns 2 and 3, Table III).

For the SMBEL function, two pairs of resonances have been interchanged as compared to the MAG order. The SMBEL matching function (eq 1) places extra emphasis on assignments where there is both a small difference between a predicted mean resonance and an observed resonance and where the range for the predicted resonance is small (i.e., it is accurately predicted). The result with this data is that the observed 50.0 and predicted mean 50.0 are matched, the observed 20.0 and predicted mean 20.0 are matched, and the observed 10.0 and predicted mean 10.0 are matched (column 4, Table IV). The remaining two matchings are then straightforward, observed 40.0 matches predicted mean 55.0, and observed 30.0 matches predicted mean 15.0.

The SDIS2 matching routine also places extra emphasis on the quality of the prediction. Predictions made at higher shell levels (there are two shell 4 predictions in this case) will tend to be assigned to nearby observed resonances. This causes the initial pairing of the observed 50.0 with predicted mean 50.0 and the observed 10.0 with the predicted mean 10.0. The remaining three assignments are as shown (column 5, Table IV). The predicted mean resonance of 22.0 (shell 2) was not matched to the observed 20.0 because making this assignment simultaneously results in a much greater error for the remaining high field resonance (i.e., observed 30.0, predicted mean 15.0).

The data for **1**, shown in Table II, constitute a more typical example of observed and predicted spectra. Identical resonance assignments result from three of the matching functions (MAG, SDIS2, and SMBEL); these assignments correspond to the ordered predicted means and observed resonances given in columns 6 and 9 in Table II. The accepted correct assignment[25] differs only with respect to an interchange of the values for C(5) and C(8). Such differences in the assignments are a result of essentially identical shifts for different resonances or are a reflection of the lack of sufficient prototypes in the data base as in this example where C(5) and C(8) were assigned as on the basis of the two-bond prototypes in the data base. Such marginal changes of assignments are not found to compromise the performance of the scoring and ranking functions. The fourth matching function, RANGE, gives a permuted set of assignments. [This matching function interchanged the assignments for the pairs C(12) and C(17), C(4) and C(6), and C(11) and C(12) and permuted the assignments for atoms C(5), C(8), C(9), and C(13)]. The RANGE assignments are rated as inferior by all the scoring functions discussed subsequently.

The evaluation of matching functions was combined with the analysis of the scoring schemes described below. This evaluation used well-characterized sets of structures and assigned spectra as discussed in the Results section.

Computing Scores Based on Spectrum Matching. Once the predicted and observed spectra have been placed in correspondence, or "matched", a score can be calculated reflecting the quality of the (mis)match. In this section we describe several different approaches to computing such a score which, like the matching functions, take into account, to varying degrees, the limitations of the spectral data and the data base. The variables required in the scoring functions described below are for the most part derived from the predicted spectrum. Each predicted resonance has a minimum, a mean, a maximum, a shell (reflecting detail of prediction), a standard deviation,[26] and the number of associated substructural models in the data base (e.g., see Table II). The scoring functions we have investigated include the following:

(A) *SUMSQ*. SUMSQ is based solely on the sum of the squares of the differences between the observed and corresponding predicted mean resonances (eq 3) (i.e., it is essentially a standard measure of Cartesian distance, taking the set of shift values as coordinates of a point in some n-dimensional space). This scoring function takes no measure of the quality of the prediction.

$$\text{SUMSQ} = \sum_{i=1}^{\text{numlines}} [\text{err}_i^2/8.0] \qquad (3)$$

where numlines is the number of observed resonances, err_i is the difference between the ith observed resonance and the ith predicted mean resonance matched to it, eight (8.0) is simply a scale factor.

(B) *SMBEL.* SMBEL uses eq 1 to compute a score for the given matching. If the SMBEL matching function was used, this score will be the "least" attainable under SMBEL matching. This scoring function takes into account the range of the predicted resonances.

(C) *SBEL.* SBEL is a measure of the belief that an observed spectrum matches a predicted spectrum. This function (eq 4) takes into account the detail of the predictions (i.e., the shell level) and the magnitude of the difference between the observed and predicted resonances

$$\mathrm{SBEL} = \sum_{i=1}^{\mathrm{numlines}} [\mathrm{shell}_i/(\mathrm{err}_i + \mathrm{default}_i)] \qquad (4)$$

where numlines is the number of observed resonances, err_i is as before, shell_i is the shell level at which the ith prediction was made, $\mathrm{default}_i = 2^{1-\mathrm{shell}_i}$ to prevent the denominator of eq 4 from going to zero. Predictions on the basis of poor models (low shell values) contribute little to this belief score. Large differences between observed resonances and matched predicted mean resonances also contribute little to the belief value.

(D) *SDIS2.* SDIS2 uses eq 3 to compute a score based on the given matching. If SDIS2 was used as a matching function, the score obtained is the "best" score under SDIS2 matching. This score is a measure of the *disbelief* that a predicted spectrum and an observed spectrum are for the same structure. A small SDIS2 value means either that there are only poor models (small *shell* values) available in the data base on which to base predictions or that the errors between the matched observed and predicted resonances are small. This accommodates limitations of the data base by discriminating *only* against candidate structures with good substructural models but large SDIS2 (disbelief) scores.

These scoring functions are evaluated in the following Results section. In all these functions, when processing incomplete observed spectra, an unmatched predicted resonance is ignored, i.e., it contributes zero to the overall score for the candidate structure.

Ranking Ordering of Candidate Structures. The final step in the procedure is to rank order the set of candidate structures on the basis of the scores obtained from the sequence of spectrum prediction, matching, and scoring. The rank ordering is trivial, based solely on the numerical values of the scores. The chemist can be discard candidate structures whose scores are judged to be so poor as to eliminate them from further consideration.

RESULTS

The evaluation of the matching and scoring functions was performed by using a method suggested by Schwenzer and Mitchell.[13] The data on alkylamines[19] were separated into a set of amines used to construct an "amine" data base[27] and a test set of hexylamines used for spectrum prediction and ranking. None of the test set was included in the data base. Because these test structures have already had their ^{13}C spectra assigned, we could evaluate not only the discriminatory power of the various functions (that is, their ability to rank the correct structure near the top of a list of possible structures) but also the accuracy of the effective assignments made by the program during the matching process.

The test consisted of treating each of the 11 available hexylamines in the test set (Table V) as an unknown. In the absence of additional structural information, the set of structural candidates for each "unknown" is the complete set of 39 $C_7H_{17}N$ amines, which we obtained from the CONGEN program.[4] For each of the 11 amines we predicted the spectra of the 39 structural candidates using the amine data base and used the four spectrum matching functions to match predicted and observed resonances (we did not make use of multiplicity information as a worst-case test of our method). For each of the matching functions, we used the four scoring functions described previously to determine the quality of the match.

Table V summarizes the results of the evaluation. Table V also summarizes the mean shell level for the predictions, derived from averaging the shell level obtained for each of the seven carbon atoms during spectrum prediction. This value is a good measure of how well each amine's substructures were represented in the data base. The mean shell varied from 1.3 (in the case of 1,2,2-trimethylpropylamine) to 3.6 (in the case of di-*n*-propylamine). Thus, substructures of the former compound were poorly represented while substructures of the latter were well represented in the data base.

The entries in the rows and columns of Table V indicate the rank order of the correct structure of the "unknown" out of the total of 39 possible structures, for the given matching and scoring functions. From Table V, it is evident that SDIS2 is the best scoring function. With only a single exception, RANGE matching for 1,2,2-trimethylpropylamine, the correct structure is always ranked first by SDIS2 independent of matching function. On the basis of this experiment and many other empirical tests of the program against known structures, we have decided that SDIS2 is the scoring function of choice. The matching functions from the data given in Table V appear to be of similar quality, with the exception of RANGE for 1,2,2-trimethylpropylamine. This also corresponds with other empirical observations. For simplicity and consistency, we have decided to use the SDIS2 function for both matching and scoring. Further support for the choice of SDIS2 scoring is presented in the next example. Indeed, the compound that was ranked most poorly in the study by Schwenzer (*N,N*-dimethyl-*sec*-butylamine ranked 6th out of 39)[13] was ranked first using SDIS2 matching and scoring.

Use of multiplicities would be expected to improve results on both the matching and the scoring processes, and indeed, the effect in this example is that all matching and all scoring routines produce the correct results (i.e., correct structure ranked first).

In a second test, again similar to a test performed by Schwenzer and Mitchell,[13] spectra were predicted for a test set of nonanes[18] on the basis of a data base of alkanes[28] (omitting the test set). In this example there are 24 nonanes available as "unknown" out of a total of 35 structural isomers of C_9H_{20} obtained from CONGEN. The procedure for prediction, matching, and scoring was as in the previous example except that only the SDIS2 matching function was used. In Table VI we present the results obtained both with and without multiplicity data.

Referring to Table VI, again excellent results are obtained even for compounds with low shell values for prediction. In all cases where the correct compound was not ranked first, use of multiplicities in the matching process increased (dramatically in some cases) the rank of the correct compound. With complete multiplicity information, the number of structural candidates is reduced significantly from the 35 possible (last column, Table VI). Again, the SDIS2 scoring function yields the best performance both with and without multiplicities. Note that the compound which yielded the poorest performance in the Schwenzer and Mitchell test (2,2,3,3-tetramethylpentane was ranked 9th out of the 35 possible nonanes) was dealt with successfully only by the SDIS2 scoring function and this in spite of the fact that the mean shell level on which predictions were based is only 1.3, i.e., on the average, only α- and some β-substituent effects are taken into consideration.

Table V. Matching and Ranking Results for the Set of Hexyl Hexylamines Based on a Library of Amines[a]

matching functions	scoring functions SUMSQ	SBEL	SDIS2	SMBEL
n-Hexylamine, Shell = 3.3				
SMBEL	1	1	1	1
RANGE	1	1	1	1
MAG	1	1	1	1
SDIS2	1	1	1	1
1,3-Dimethylbutylamine, Shell = 2.3				
	1	3	1	2
	1	3	1	2
	1	3	1	2
	1	3	1	2
1,2,2-Trimethylpropylamine, Shell = 1.3				
	9	11	1	1
	34	16	21	5
	12	9	1	1
	11	10	1	1
2,2-Dimethylbutylamine, Shell = 1.8				
	1	7	1	2
	1	6	1	2
	1	7	1	2
	1	7	1	2
Di-*n*-propylamine, Shell = 3.6				
	1	1	1	1
	1	1	1	1
	1	1	1	1
	1	1	1	1
Diisopropylamine, Shell = 2.6				
	1	2	1	1
	1	2	1	1
	1	2	1	1
	1	2	1	1
N-Ethylbutylamine, Shell = 3.5				
	1	1	1	1
	1	1	1	1
	1	1	1	1
	1	1	1	1
N-Ethyl-*sec*-butylamine, Shell = 2.5				
	1	1	1	1
	1	1	1	1
	1	1	1	1
	1	1	1	1
Triethylamine, Shell = 2.5				
	1	1	1	1
	1	1	1	1
	1	1	1	1
	1	1	1	1
N,*N*-Dimethyl-*sec*-butylamine, Shell = 1.6				
	1	8	1	3
	1	6	1	3
	1	6	1	3
	1	6	1	3
N,*N*-Dimethyl-*tert*-butylamine, Shell = 1.5				
	1	14	1	1
	1	13	1	2
	1	13	1	2
	1	13	1	2

[a] For each amine the table indicates results obtained by using all four matching with all four scoring functions. The numbers in the table are the ranking of the correct structure out of the 39 possible hexylamines.

Table VI. Ranking of 24 Nonanes Analyzed as Unknowns[a]

compound	shell	SUMSQ	SBEL	SDIS2	SMBEL	out of[b]
n-nonane	3.8	1,1	1,1	1,1	1,1	1
2-methyloctane	3.5	1,1	1,1	1,1	1,1	5
3-methyloctane	3.3	1,1	1,1	1,1	1,1	5
4-methyloctane	3.2	1,1	1,1	1,1	1,1	5
2,3-dimethylheptane	2.8	2,1	2,1	2,1	2,1	9
2,4-dimethylheptane	2.7	2,1	1,1	2,1	1,1	9
2,5-dimethylheptane	2.8	1,1	1,1	1,1	1,1	9
2,6-dimethylheptane	3.5	1,1	1,1	1,1	1,1	9
3,4-dimethylheptane	2.6	1,1	1,1	1,1	1,1	9
3,5-dimethylheptane	2.7	1,1	1,1	1,1	1,1	9
2,2-dimethylheptane	3.0	1,1	1,1	1,1	1,1	5
3,3-dimethylheptane	2.7	1,1	1,1	1,1	1,1	5
4,4-dimethylheptane	2.5	1,1	1,1	1,1	1,1	5
2,3,5-trimethylhexane	2.2	2,1	5,1	1,1	2,1	3
2,2,4-trimethylhexane	2.2	1,1	1,1	1,1	2,1	8
2,2,5-trimethylhexane	2.3	1,1	4,1	1,1	1,1	8
2,3,3-trimethylhexane	2.1	1,1	2,1	1,1	3,1	8
2,2,3,4-tetramethylpentane	1.4	8,1	24,2	5,1	6,1	2
2,3,3,4-tetramethylpentane	1.3	23,1	3,1	1,1	4,1	2
2,2,3,3-tetramethylpentane	1.3	24,1	10,1	1,1	5,1	2
3-ethylheptane	3.2	1,1	1,1	1,1	1,1	5
3-ethyl-2,4-dimethylpentane	1.4	8,1	16,3	1,1	6,3	3
3,3-diethylpentane	1.3	1,1	15,3	1,1	2,1	5
2,2,4,4-tetramethylpentane	2.5	1,1	1,1	1,1	1,1	2

[a] The SDIS2 matching function was used throughout. For each alkane the mean shell for prediction is reported, along with the rank order of the correct structure, without and with multiplicities, for each of the four ranking functions. When multiplicities are not used, the rank is out of 35 possible nonanes. [b] When multiplicities are used the rank is out of the number given in the final column.

In the final example, intended to be more typical of real structural applications, we used available data for a pregnane derivative.[29] The derivative was known to incorporate a pregnane skeleton **2**. From the ^{1}H and ^{13}C NMR spectra the following substructural information could be inferred. Two methyl groups appeared as singlets in the ^{1}H NMR spectrum [C(18) and C(19)] and one appeared as a doublet [C(21)]. Three secondary acetoxy groups were present, one of which must be placed at C(20) to correspond with the observed doublet methyl C(21). Given these structural constraints, the GENOA isomer generating program[5] produced 45 constitutional isomers which differ only in the substitution pattern of the remaining two acetoxy groups about the skeleton of **2**. Spectra were predicted for each of these isomers and matched and scored with the observed spectrum by using only the *SDIS2* functions. The distribution of scores is given in Table VII under various assumptions on the quality of available spectral data. (Only the top 28 structures are included in Table VII because after the 28th there is a significant break in the distribution of scores and the remaining 17 candidates can be safely rejected.)

Along with results based on a complete observed spectrum are results based on four major types of incomplete spectra (as described previously). As one moves from left to right across Table VII, the degree of incompleteness of the spectrum increases. The effect on the rank ordering is that, although candidate 39 remains top ranked, other rankings are shuffled, particularly among closely ranked structures in columns 2 and 3. Most significantly, the discriminatory power of the ranking decreases in that the total range of scores decreases as does the interval between scores. This is most apparent, in this example, comparing columns 2 and 3 (complete spectrum with multiplicities) with columns 10 and 11 (no multiplicities, missing singlets).

Included at the bottom of each of the five groups of results is a relative timing factor. There is a dramatic increase in the computational cost (nearly two orders of magnitude) if one is forced to deal with spectral data with no multiplicities and no singlet resonances. Thus, it is important to have as much

Table VII. Ranking and Scoring Results for the Top Ranked 28 of the 45 Possible Triacetoxypregnanes[a]

	all resonances + multiplicities		missing two singlets		even/odd parity only		no multiplicities		no multiplicities, missing two singlets	
rank	compd[b]	score	compd	score	compd	score	compd	score	compd	score
1	39	11.9	39	11.8	39	11.9	39	11.9	39	11.8
2	37	18.8	26	17.5	38	18.4	38	18.4	38	13.1
3	26	19.6	37	18.0	37	18.7	37	18.7	28	15.3
4	38	19.7	38	18.9	26	20.8	44	20.1	26	16.3
5	5	26.1	28	24.7	5	26.1	26	20.5	37	17.5
6	28	26.8	5	25.0	44	26.6	28	22.1	44	18.0
7	44	27.6	13	27.1	13	28.2	7	23.6	36	20.7
8	13	28.2	44	27.6	43	28.7	29	24.7	7	21.7
9	43	30.6	43	29.7	29	29.9	13	25.7	43	21.9
10	7	31.1	7	30.0	7	30.1	5	26.1	29	22.2
11	15	31.9	15	30.8	28	30.9	43	26.6	13	22.3
12	1	36.4	29	34.4	1	31.1	1	26.7	1	22.4
13	29	36.6	1	35.6	15	31.2	25	28.7	15	22.9
14	11	37.8	11	36.4	11	36.9	15	28.9	5	23.4
15	3	38.8	36	37.1	3	37.8	11	30.6	40	26.5
16	25	44.7	3	37.5	8	40.8	8	33.4	25	27.3
17	8	45.5	40	39.4	36	40.8	33	33.8	33	27.4
18	36	45.8	25	43.3	40	41.1	40	33.9	11	29.0
19	16	47.1	8	44.1	16	42.5	3	34.1	32	29.3
20	30	47.5	16	45.8	25	45.7	30	36.4	3	29.4
21	40	48.1	30	46.1	30	46.4	45	36.8	8	31.3
22	45	48.7	45	48.5	45	48.1	16	38.2	22	31.8
23	33	49.7	33	49.6	33	49.4	36	39.8	27	31.9
24	27	54.2	27	53.4	27	50.6	32	43.0	20	33.0
25	14	62.3	14	58.7	14	54.9	42	43.9	30	33.8
26	6	63.0	6	59.4	6	58.0	27	46.8	45	34.5
27	17	65.8	41	60.5	41	61.8	22	47.7	16	35.3
28	41	69.2	17	64.3	17	64.8	6	48.3	31	37.6
Break[c]	5.8		6.7		12.4		3.6		1.5	
Time[d]	1.0		1.0		19		62		85	

[a] Column 1 gives the rank order. The next five pairs of columns are results based on the most common forms of observed spectra. The pairs of columns give the results when (1) all resonances and multiplicities are known, (2) the two sp^3 singlets, C(10) and C(13), are missing, (3) only the parity of the resonances is available, (4) no multiplicities were determined, and (5) no multiplicities available and the two singlets missing. [b] The number of the compound in the set of 45 candidate structures. [c] The difference in score between the 28th and 29th candidate structure. [d] The relative amount of computer time required to obtain the rankings.

multiplicity information available as possible not only for the obvious increase in discriminatory power but also to decrease computational costs.

This particular "unknown" is of sufficient structural complexity that spectral prediction including stereochemistry is warranted. The top 11 candidates were selected for further study on the basis of the break in the distribution of scores (column 3, Table VII) derived from the complete spectrum plus multiplicities. Stereoisomers were generated for these 11 constitutional isomers,[9] using as constraints that the unknown must possess the configurations of the standard 5α-pregnane skeleton (**2**) and that the C(20) acetoxy is in the *R* configuration (assumed by the original authors[29]). Since the only configurations not designated were those of the two remaining acetoxy-bearing carbons, each constitutional isomer gave rise to four stereoisomers. Matching of the observed spectrum to the predicted spectra of these isomers, now considering stereochemistry, followed by subsequent scoring and ranking gave the results shown in Table VIII. The isomer ranked highest based on consideration only of molecular constitution, No. 39 (Table VII) yielded the four highest ranking stereoisomers. The highest ranked stereoisomer, 3β,6α,20(*R*)-triacetoxy-5α-pregnane (**3**) is in fact the structure assigned to the unknown by the original authors.[29]

DISCUSSION

We have presented our approach to use of predicted ^{13}C NMR spectra as an aid to evaluation of candidate structures for an unknown compound. We also have presented results from experiments on the relative merits of different methods for matching predicted and observed spectra and subsequent derivation of a score which reflects the quality of the matching. The prediction process itself is conceptually straightforward, given that the data base is structured to contain relevant information on shift ranges and distributions and shell level descriptions of substructural environments.

The matching routine of choice is the SDIS2 function, primarily because we purposefully designed it to reflect the limitations of the data base. It is gratifying that it also performs well in cases where observed spectra lack multiplicity information and/or have missing resonances. The SDIS2 scoring function also yields the best performance; thus we have the advantage of both optimizing the match and calculating a "best" score in a single step.

The SDIS2 functions have proven satisfactory in that they allow for distinction between structures that are well represented by substructures in the data base without discriminating against novel systems for which only poor (low shell level) predictions can be made. Thus, it is possible to reduce the number of candidates, for an unknown, to a few likely structures and a set of unusual structures retained because of poor substructural representation in the data base. Frequently, such unusual structures can be eliminated on the basis of other data or, for natural products, through consideration of biosynthetic constraints. As the data base is expanded, a continuing process, predicted spectra will be of higher quality because it will be less common to retain structural candidates simply because of lack of appropriate substructures in the data base.

Table VIII. Results of Ranking the Stereoisomers of the Top Eleven Ranked Compounds [Table VII, Columns 1 and 2 (Using All Resonances and Multiplicities)][a]

	acetate substituent		rank order of candidate structure	
compd[b]	location	(config)	without stereochemistry[b]	with stereochemistry[c]
39	3 (β)	6 (α)	1	1
	(α)	(α)		2
	(β)	(β)		3
	(α)	(β)		4
37	4 (β)	6 (α)	2	9
	(α)	(α)		9
	(β)	(β)		13
	(α)	(β)		13
26	6 (β)	11 (α)	3	6
	(α)	(α)		7
	(β)	(β)		11
	(α)	(β)		11
38	2 (α)	6 (α)	4	5
	(β)	(α)		8
	(α)	(β)		15
	(β)	(β)		16
5	6 (α)	16 (β)	5	22
	(β)	(β)		17
	(α)	(α)		27
	(β)	(α)		21
28	4 (β)	11 (β)	6	18
	(α)	(β)		18
	(β)	(α)		23
	(α)	(α)		23
44	3 (β)	4 (β)	7	30
	(α)	(β)		30
	(β)	(α)		30
	(α)	(α)		30
13	6 (α)	15 (α)	8	34
	(β)	(α)		20
	(α)	(β)		36
	(β)	(β)		29
43	4 (β)	16 (β)	9	40
	(α)	(β)		39
	(β)	(α)		38
	(α)	(α)		35
7	2 (α)	4 (β)	10	41
	(β)	(β)		41
	(α)	(α)		37
	(β)	(α)		37
15	4 (β)	15 (α)	11	25
	(α)	(α)		26
	(β)	(β)		43
	(α)	(β)		44

[a] The position of substitution of the two nuclear acetoxy groups is based on the standard pregnane numbering (2). The αs and βs designate the stereochemistry of the acetoxy group at the indicated position. [b] Compound number and rank order obtained from column 1 and 2, Table VII. [c] Rank order of indicated stereoisomer in the set of 44 possibilities (note that tie scores are given the same rank order).

In the results presented, we were able to identify the correct structure for all unknowns. We are not so sanguine as to presume that this will always be the case. Our method for spectrum prediction and structure ranking should be viewed as a procedure which can narrow down the possibilities for an unknown structure; there is no guarantee that the top-ranked structure will be the correct structure. Rather, we use this technique as a method of rejecting those structural candidates which are clearly unreasonable on the basis of the lack of agrement between predicted and observed spectra, leaving it for the chemist and additional experimental data to discriminate among the remaining possibilities.

One limitation of our approach structure ranking scheme is that there must not be an excessive number of candidate structures. Typical problems analyzed with the aid of our structure generator programs[4,5] yield from a few dozen up to perhaps three or four hundred candidate structural isomers. The ^{13}C analysis program can handle problems of such dimension. However, the scope of such a problem is greatly expanded when stereoisomerism is considered. Usually, it is then necessary to solve the problem in stages, as with the pregnane derivative **3**. First, the ^{13}C analysis is performed on constitutional isomers to attempt to eliminate as many as possible. Subsequently, a constrained stereoisomer generation procedure can be applied and the predicted ^{13}C spectra of the stereoisomers further analyzed.

Another limitation is the inadequacy of the data base used for prediction. Although the matching and scoring functions can accommodate missing data, there is no way to correct for erroneously assigned reference spectra or reference spectra showing solvent, concentration, or temperature dependent effects; these factors result in broadened shift ranges for particular substructures. Such broadening reduces the discriminatory power of our approach. We attempt to control these factors by rigorous checks at the time of data entry into the data base, together with periodic overall checks of the consistency of data in the data base. Such problems are of much less significance in applications using specialized libraries of spectral data from similar compounds run under standardized conditions.

EXPERIMENTAL SECTION

These programs are implemented in the ALGOL-like BCPL program language[30] on a Digital Equipment Corporation KI-10 computer at SUMEX-AIM 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, or other computers supporting BCPL (e.g., IBM-370), is possible. However, additional work remains before the programs become polished enough for mass export. In the meanwhile, within the limits of our resources, we are prepared to collaborate in the ^{13}C-based solution of nontrivial structure problems for outside investigators who lack appropriate computer facilities.

ACKNOWLEDGMENT

We wish to thank the National Institutes of Health (RR-00612) for their generous financial support. Computer resources were provided by the SUMEX facility at Stanford University under National Institutes of Health Grant RR-0785.

REFERENCES AND NOTES

(1) Part 39 of the series "Applications of Artificial Intelligence for Chemical Inference". For part 38, see D. H. Smith, N. A. B. Gray, J. G. Nourse, and C. W. Crandell, *Anal. Chim. Acta, Comput. Tech. Optim.*, **133**, 471 (1981).

(2) C. A. Shelley, T. R. Hayes, M. E. Munk, and R. V. Roman, *Anal. Chim. Acta* **103**, 121 (1978).

(3) V. V. Serov, M. E. Elyashberg, and L. A. Gribov, *J. Mol. Struct.*, **31**, 381 (1976).

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

(5) R. E. carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse, and C. Djerassi, *J. Org. Chem.*, **46**, 1708 (1981).

(6) S. Sasaki, H. Abe, Y. Hirota, Y. Ishida, Y. Kudo, S. Ochiai, K. Saito, and T. Yamasaki, *J. Chem. Inf. Comput. Sci.*, **18**, 211 (1978).

(7) L. A. Gribov and M. E. Elyashberg, *CRC Crit. Rev. Anal. Chem.*, **18**, 110 (1979).

(8) Z. Hippe and R. Hippe, *Appl. Spectrosc. Rev.*, **16**, 135 (1980).

(9) J. G. Nourse, D. H. Smith, and C. Djerassi, *J. Am. Chem. Soc.*, **102**, 6289 (1980).

(10) L. A. Gribov, M. E. Elyashberg, and M. M. Raikhshtat, *J. Mol. Struct.*, **53**, 81 (1979).

(11) N. A. B. Gray, C. W. Crandell, J. G. Nourse, D. H. Smith, M. L. Dageforde, and C. Djerassi, *J. Org. Chem.*, **46**, 703 (1981).

(12) N. A. B. Gray, J. G. Nourse, C. W. Crandell, D. H. Smith, and C. Djerassi, *Org. Magn. Reson.*, **15**, 375 (1981).

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

(14) 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).

(15) 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).

(16) L. A. Gribov, M. E. Elyashberg, and V. V. Serov, *J. Mol. Struct.*, **50**, 371 (1978).

(17) H. B. Woodruff, C. R. Snelling, Jr., C. A. Shelly, and M. E. Munk, *Anal. Chem.*, **49**, 2075 (1977).

(18) L. P. Lindeman and J. Q. Adams, *Anal. Chem.*, **43**, 1245 (1971).

(19) H. Eggert and C. Djerassi, *J. Am. Chem. Soc.*, **95**, 3710 (1973).

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

(21) H. Hambloch and A. W. Fralm, *Tetrahedron*, **36**, 3273 (1980).

(22) G. A. Morris and R. Freeman, *J. Am. Chem. Soc.*, **101**, 760 (1979).

(23) G. E. Martin, J. A. Matson, J. C. Turley, and A. J. Weinheimer, *J. Am. Chem. Soc.*, **101**, 1888 (1979).

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

(25) C. L. VanAntwerp, Ph.D. Thesis, Stanford University, Stanford, CA, 1978.

(26) The standard deviation of the observed shift distributions is not used in any scoring function since many distributions of resonances are non-Gaussian. In fact, bimodal distributions are common and can be caused by the presence of different configurational forms. For example, the predicted range for C(19) of **1** (see Table II) is wide because it was based on the 5α and 5β forms of the steroid nucleus.

(27) The library contains the following structures and spectra: pentylamine, 1-methylbutylamine, 2-methylbutylamine, 3-methylbutylamine, 2,2-dimethylpropylamine, *N*-methyl-*sec*-butylamine, *N*-methyl-*tert*-butylamine, *N*-methyldiethylamine, heptylamine, 1-methylhexylamine, 1-ethylpentylamine, 1,3-dimethylpentylamine, *N*-methylhexylamine, *N*-ethylpentylamine, *N*-isopropylbutylamine, *N*-isopropyl-*sec*-butylamine, octylamine, 1-methylheptylamine, 2-ethylheptylamine, 1,5-dimethylhexylamine, 1,1,3,3-tetramethylbutylamine, dibutylamine, diisobutylamine, *N*-ethylhexylamine, *N,N*-dimethylhexylamine, *N,N*-diethylbutylamine, *N,N*-diethyl-*sec*-butylamine, *N*-ethyldiisopropylamine, nonylamine, 1-isopropylhexylamine, *n*-propylhexylamine, *N*-*sec*-butylpentylamine, *N*-*sec*-butyl-3-methylbutylamine, *N*-*tert*-butyl-3-methylbutylamine, *N*-methyl-1,1,3,3-tetramethylbutylamine, tripropylamine, decylamine, dipentylamine, *N*-butylhexylamine, *N*-*tert*-butylhexylamine, *N*-*sec*-butyl-3,3-dimethylbutylamine, bis(3-methylbutyl)amine, *N*-ethyldibutylamine, *N,N*-diisopropylbutylamine, *N*-pentylhexylamine, *N*-butyl-1-methylhexylamine, *N*-pentyl-1,3-dimethylbutylamine, *N*-pentyl-1,2,2-trimethylpropylamine, *N*-(3,3-dimethylbutyl)pentylamine, dihexylamine, *N*-butyl-1-ethylpentylamine, *N*-methyl-*N*-butylhexylamine, *N*-propyldibutylamine, *N*-isopropyldibutylamine, *N*-(1,3-dimethylbutyl)hexylamine, tributylamine, *N*-ethyldipentylamine, *N*-*tert*-butyldibutylamine, *N*-pentyl-1,1,3,3-tetramethylbutylamine, *N,N*-dibutyl-3-methylbutylamine, *N*-(1-ethylpentyl)-1-propylbutylamine, *N,N*-dibutylhexylamine, *N,N*-dibutyl-3,3-dimethylbutylamine, *N*-*sec*-butyldipentylamine, and *N,N*-dibutyl-1-methylpentylamine.

(28) The library contains the following structures and spectra: pentane, 2-methylbutane, 2,2-dimethylbutane, hexane, 2-methylpentane, 3-methylpentane, 2,2-dimethylbutane, 2,3-dimethylbutane, heptane, 2-methylhexane, 3-methylhexane, 2,3-dimethylpentane, 2,4-dimethylpentane, 2,2-dimethylpentane, 3,3-dimethylpentane, 2,2,3-trimethylbutane, 3-ethylpentane, octane, 2-methylheptane, 3-methylheptane, 4-methylheptane, 2,3-dimethylhexane, 2,4-dimethylhexane, 2,5-dimethylhexane, 3,4-dimethylhexane, 2,2-dimethylhexane, 3,3-dimethylhexane, 2,3,4-trimethylpentane, 2,2,3-trimethylpentane, 2,3,3-trimethylpentane, 2,2,4-trimethylpentane, 2,2,3,3-tetramethylbutane, 3-ethylhexane, 3-ethyl-2-methylpentane, and 3-ethyl-3-methylpentane.

(29) J. W. ApSimon, S. Badripersaud, J. A. Buccini, J. Enkhoorn, and M. W. Gilgan, *Can. J. Chem.*, **58**, 2703 (1980).

(30) M. Richards and C. Whitby-Strevens, "BCPL—the Language and its Compiler", Cambridge University Press, Cambridge, 1979.