Computer Aided Method for Chemical Structure Elucidation Using Spectral Databases and ^{13}C NMR Correlation Tables

Igor I. Strokov and Konstantin S. Lebedev*

Novosibirsk Institute of Organic Chemistry, Siberian Branch of Russian Academy of Science, Lavrentiev Avenue 9, Novosibirsk 90, Russia

Received August 31, 1998

The work describes a new computer method for structure elucidation using spectral databases and ^{13}C NMR correlation tables. Databases provide retrieval of compounds with spectra closest to the query, whereas correlation tables find application on the three steps: derivation of structural fragments consistent with the experimental ^{13}C NMR spectrum, generation of structures with a molecular formula and a set of derived fragments, and ranking of generated structures by the similarity of their predicted ^{13}C NMR spectra to the experimental one. As it is shown in the examples of both test and real problems, the approach provides revelation of large structural fragments and generation of a small number of candidate structures.

1. INTRODUCTION

The analysis of spectral data for the unknown compound structure determination remains a usual and laborious task in chemical practice. Since the advent of computers much efforts has been directed toward facilitating the solution of this problem.[1,2] Two trends are recognized in the computer-assisted structure elucidation. The first one, originated from modeling a human way of reasoning, has led to artificial intelligence (AI) systems, that use formalized knowledge bases to store both substructure-subspectra correlations and the rules of using this information. The knowledge bases contents define the general scope of problems that can be solved with the help of AI systems.[3−8]

The other approach assumes primary use of large collections of molecular spectra.[9−13] Compared to knowledge bases, spectral databases are usually more common and involve a larger and faster replenishing amount of data. Besides, searching databases allows quick identification of previously registered compounds. However, the methods for structure elucidation of compounds outside a database are not exhaustively studied, especially when combining various kinds of molecular spectra. At the same time, continuation of efforts in this direction is quite attractive, promising to integrate AI advantages with those provided by databases use. Following this direction we suggest a new method for structure elucidation using mass and ^{13}C NMR spectra databases and ^{13}C NMR correlation tables (CT).

2. METHODS

Let us skip the fortunate case when an unknown compound is found in a database. Instead of an unknown compound itself, a close structure, which has a spectrum similar to the query, is commonly found. The following example illustrates this situation.

Let compound X_1 be obtained from a synthetic experiment,[14] and let a mass spectral database contain, instead of

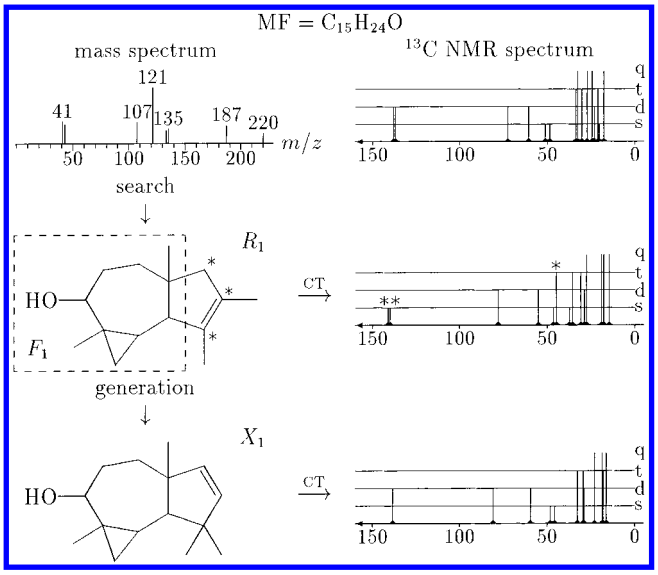


Figure 1. The structure elucidation scheme on the example of compound X_1. Heights in ^{13}C NMR spectra reflect peaks multiplicity. Asterisks mark unmatched signals and corresponding carbon vertices.

X_1, its structural isomer R_1, which is found in a spectral search result (Figure 1). The right decision seems to be nearby: if one could derive a *correct* (i.e., possessed by X_1) fragment from R_1, then the remainder of the molecule could be easily built by means of a structure generator.

How could a correct fragment be recognized? The statistical method (which may be originated from the work[15]) is based on the assumption that the relative occurrence of correct fragments should be higher in the spectral search result (SSR) compounds than in the entire database. This method works better for more abundant fragments, which evidently do not encompass F_1. Besides, a statistical evaluation often has a low reliability. As shown in ref 16, even a mutual use of SSR on different kinds of spectra cannot insure against faults.

* Author to whom inquiries should be sent. E-mail: leb@nioch.nsc.ru.

10.1021/ci980184p CCC: $18.00
Published on Web 05/18/1999

Application of spectral data for fragments determination could certainly improve the results. Because of their ability to assign a spectral signal to a carbon atom, ^{13}C NMR spectra are perfectly suited for this task. Let us assume that the one-dimensional ^{13}C NMR spectrum of an unknown compound (spectrum *X*) is available. Then the main idea of the proposed method is very simple: for all SSR compounds we predict ^{13}C NMR spectra and find all the fragments whose subspectra do not conflict with spectrum *X*. Further a common scheme is used: the fragments are sent to a structural generator, and then generated structures are ranked due to their predicted ^{13}C NMR spectra similarity to spectrum *X* (see Figure 1).

3. DERIVATION OF FRAGMENTS

Let us return to "unknown" compound X_1. Its mass, ^{13}C NMR spectra, and molecular formula (MF) are given (the problem of MF recognition with the aid of low resolution mass and optionally ^{13}C NMR and ^{1}H NMR spectra is discussed in the paper[17]). The library search on the mass spectrum results in the single close analog R_1, while SSR on the ^{13}C NMR spectrum contains no similar structures at all. This remark can single out the feature of our approach to derive fragments from any structures, not confining their set to a ^{13}C NMR database (cf. refs 13 and 18−20).

Anyhow, in our case compound R_1 turns out to be sufficient to solve the problem. After predicting the ^{13}C NMR spectrum for R_1 (the used method is discussed below) three predicted signals related to three five-membered ring vertices (marked by an asterisk on the figure) occur to mismatch any experimental signal. After rejecting these vertices along with the adjacent methyl groups there remains fragment F_1 which allows building nine molecular graphs with the given MF. Ranking by means of ^{13}C NMR prediction and comparison with spectrum *X* brings the correct structure to the first place.

The derivation of fragments is a quite complex and multivariate process indeed. Many other fragments are obtained in addition to F_1. Processing each structure in SSR, the algorithm tries to mark some vertices (a set of marked vertices spawns a fragment) by all ways so that a set of signals (a subspectrum) corresponding to marked vertices would not conflict with spectrum *X*. This condition is true if any signal from a subspectrum can be related to one signal from *X*, and vice versa, within the same *group* of signals.

In order to explain *groups*, let us define what is *signal matching* first. Predicted signal *r* and experimental signal *x* match if they have the same multiplicity ($m_r = m_x$) and close chemical shifts

$$|\delta_r - \delta_x| \le \Delta \qquad (1)$$

(signals belonging to the same spectrum do not match by definition). Let us use bilateral graphs to represent signal matching: signals (vertices) are connected if and only if they match (Figure 2). Then *groups* are connected subgraphs of the matching graph.

In the algorithm groups are considered independently from each other. Also, processing of a group depends on its kind. There are three kinds of groups.

First, note "alone" signals those match no one. Vertices in structure *R* related to alone signals *r* are never marked and cannot fall into the extracted fragments.

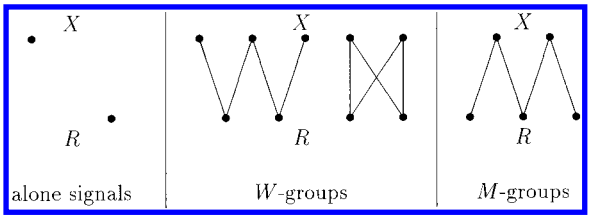
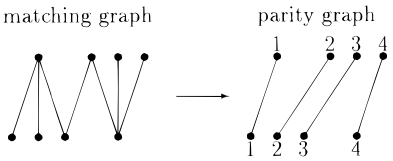


Figure 2. Signal groups in compared spectra *R* (predicted) and *X* (experimental).

Then consider *W*-groups (the notation comes from the view of subgraphs) which contain *no more* signals *r* than *x*. Vertices related to signals *r* in *W*-groups are *always* marked, since a unique *x* pair can be assigned to any signal *r*. To find (*r*, *x*) pairs we simply rank *r* and independently *x* signals by δ values and then combine the signals of same ordinal numbers. We realize that this method can lead to unmatching pair signals, as is shown for pairs 2−2 and 3−3 below:



Thus belonging to one group turns out to be a softer requirement than the signal matching. This rebate allows both to simplify the marking algorithm and sometimes to derive larger fragments.

In both cases all vertices *r* are either rejected (alone signals) or marked in advance (*W*-groups). The intermediate position is occupied by *M*-groups (also called so because of their shape) which have *more r* signals than *x*. Let the *M*-group have *m* signals *r* (and *m* corresponding vertices) and *n* signals *x*, where $m > n$. Evidently, only *n* vertices from *m* can be marked by C_m^n ways. After choosing the next combination of *n* from *m*, signals are broken into pairs in the same way as in *W*-groups.

Thus the enumeration touches only *M*-groups. If several *M*-groups exist, then all joined combinations are considered. If their number is above a given limit, then one of two restrictive means is used: to consider only the combinations providing the least sum difference of chemical shifts for pair signals or to refuse marking vertices in *M*-groups at all.

Signals matching and grouping primarily depend on the admissible difference of chemical shifts denoted in (1) as Δ. Its value giving the maximal correct (optimal) fragments may differ for various tasks and analyzed structures. The algorithm operates with a Δ range instead of a fixed value lest the optimal solution is missed. In the program the given range is broken into several sections so that within a section signals grouping remains intact. As all further marking of vertices depends only on the grouping, it is enough to take a single Δ value from each section.

Thus, the analysis of a structure results in a set of fragments. Each carbon atom in a fragment has a unique assignment to some signal in spectrum *X*. Heteroatoms give no signals in ^{13}C NMR spectra. However, they influence on δ of neighboring carbons and therefore can be confirmed by implication. A heteroatom is marked in two cases: the number of adjacent marked carbons exceeds the number of nonnmarked ones or (the second case) all its neighbors are

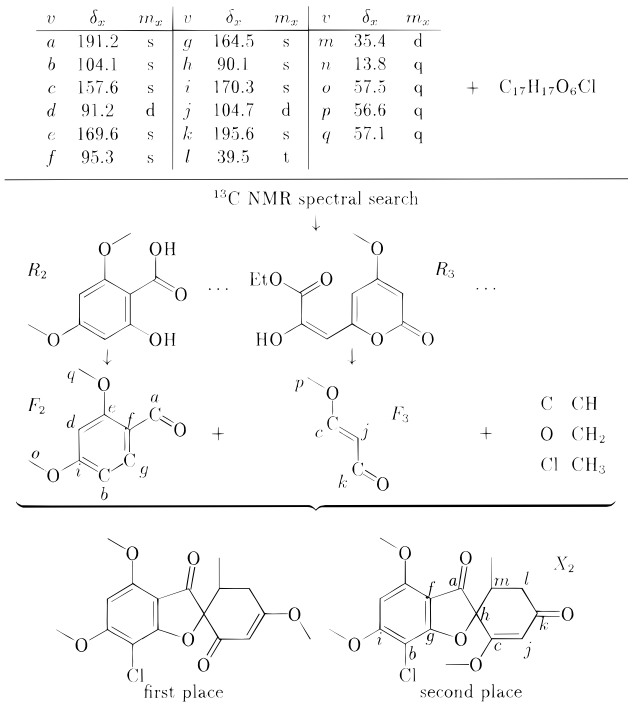


Figure 3. Elucidation of structure X_2 by its ^{13}C NMR spectrum.

marked. It means, for example, that in group NO_2 adjacent to a marked carbon all three heteroatoms will be marked too.

4. USING FRAGMENTS

Above we considered the main principles of extracting the fragments which fit ^{13}C NMR spectrum *X*. However, some important details are still worth consideration by the following example.

Let only MF and ^{13}C NMR spectrum of "unknown" compound X_2 be known (see Figure 3). The available ^{13}C NMR database contains no close analogs of X_2 except for R_2 and R_3, found respectively at the fourth and 15th places of the SSR. However, even these structures would not help without the following specific techniques.

Fitting SSR Structures to Spectrum *X*. Consider, for example, structure R_2. Its predicted ^{13}C NMR spectrum (which surely almost completely coincides with the experimental one from the database) has four singlets (96.4, 161.4, 165.1, 171.9 ppm) and two doublets (90.9 and 93.9) related to the aromatic ring carbons. At the same time spectrum X_2 contains only one doublet (91.2) in this region. If the above rules are strictly followed, then only one aromatic vertex −CH═ from the two could be marked. It means a disadvantageous ring disclosure makes the generator build many acyclic structures. Meantime, the existence of a 5-substituted aromatic ring is quite evident. In this case spare −CH═, instead of being rejected, is replaced by >C═ and then marked along with the other ring vertices. The replaced vertex is assigned an artificial singlet with a wide range of δ corresponding to a substituted aromatic carbon. As a result, the structure R_2 will produce, among other fragments, fragment F_2, which, strictly speaking, does not belong to R_2. Thus fitting to spectrum *X* consists in preserving of aromatic rings by means of introduction or removal of dummy substituents.

Combining Fragments. Fragment F_2 composes a significant part of compound X_2. It is, however, insufficient to build a manageable number (say, < 10 000) of structural hypotheses. Meanwhile, structure R_3 produces fragment F_3, which represents *another* part of X_2 (see Figure 3). In other words, F_2 and F_3 do not overlap. It can be ascertained with the aid of available carbon-signal bindings without the knowledge of structure X_2. In our case F_2 vertices refer to signals {*a,b,d,e,f,g,i,o,q*}, whereas F_3—to signals {*c,j,k,p*}. As these subsets do not intersect, both F_2 and F_3 can be used in the generation step simultaneously (as one nonconnected fragment).

Initial data commonly include MF, ^{13}C NMR spectrum of unknown *X*, and SSR. The program analyzes SSR structures in turn to derive all different fragments from them. Each derived fragment is stored and combined with each previously stored one. The combination implies that two fragments are considered as one nonconnected fragment which is then made to conform spectrum *X* by deleting, if necessary, some vertices. The conformity, in turn, means that the sum of references to signals groups should not exceed the number of signals in these groups.

Spectral Bans on Bond Formation. After the next combination of two fragments (denote it F_{i+j}) is obtained, an attempt to generate all molecular structures with F_{i+j} is made. If the generator builds not more than n_{max} structures, then they are all stored for the subsequent analysis. Otherwise the generation is interrupted at $n_{max} + 1$ structure without retaining the partial result. Parameter n_{max} was called above "the manageable number of hypotheses", i.e., treatment of that number of structures should not exceed computer resources and the user's patience. In most cases in our practice n_{max} < 1000 was enough to solve a task within a few minutes.

In task X_2 the combination F_{2+3} does lead to the correct solution (see Figure 3). The multiplicities of signals *h*, *l*, *m*, *n*, unreferenced by F_{2+3} carbons, testify the existence of four vertices C, CH, CH_2, CH_3 in X_2. Subtraction of atoms constituting these vertices and fragment F_{2+3} from the formula $C_{17}H_{17}O_6Cl$ leaves one chlorine and one oxygen atom, which act as distinct vertices in the generation. All the distinct vertices (six ones altogether) and some of the fragment vertices (C_a, C_g, C_b, C_c, C_k) have free bonds. Binding these bonds together by all possible ways essentially constitutes the generator function. Its scope, however, can be restricted by setting the maximal bond order for every pair of nonequivalent vertices.[21,22]

In this example 27 168 structures are built without bond order limitations. Among them many structures have oxygen atom bound with vertex C_b instead of C_g as it should be. At the same time bond C_b−O can be forbidden in advance (by setting its maximal order to 0) taking into account that vertex C_b refers to signal *b* with δ = 104.1 ppm. Meantime, aromatic vertex C adjacent to the oxygen atom usually has a higher δ: the correlation tables give the range 124.7−184.7 with the mean value of 145.3. Thus, if C_b would be bound with oxygen, its δ increased by ≃40 ppm. As this difference surpasses prediction inaccuracy, the bond could be forbidden, because the references to signals are presumed to persist regardless of which structure a fragment belongs to. In other words, the signal predicted for a vertex after the change of its environment should not differ from the

experimental signal assigned to this vertex more than for a given value.

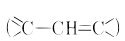
Though not all bonds can be checked this way (at least one vertex of a bond is to contain a carbon atom whose environment should be known in a distance of one or more bonds after the bond creation), involving the spectral information before the generation step can drastically reduce both generation time and the number of resulting structures. In the example with fragment F_{2+3} 223 structures are generated instead of 27 168. The overall statistics on task X_2 include 387 nonisomorphic structures ($n_{max} = 700$) which, after comparing their predicted ^{13}C NMR spectra with spectrum X_2, include the correct structure at the second place of the ranked list of candidates (Figure 3).

5. PREDICTION OF ^{13}C NMR SPECTRA

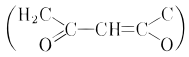
The first place in the ranked list is still occupied by the isomer of unknown compound X_2. This fact is indicative of insufficient prediction accuracy. It is not too surprising because the database, used to compose the CT, contains no close structural analogs of X_2.

The present CT are maintained by the index-sequential access method (ISAM) implementation in the file system JoKey[25] which is characterized by the perfect performance. Single records contain statistical data (mean, maximal, and minimal δ, etc.) on a carbon atom in a particular environment. The wide treelike code[23] of an atom and all its neighbors at the distance of one, two, or three bonds composes a record key. This approah still originates from Bremser's HOSE/HORD codes.[26] This way excels in simplicity, small storage requirements (in our case CT built on the basis of 26 000 molecular structures and spectra occupy 966 656 bytes), and high prediction rates due to the direct access use. On the other hand, only fully described environments can be accessed directly. If the exact match is absent, then a lower environment level has to be tried.

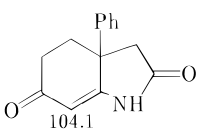
Consider, for example, prediction of the ^{13}C NMR spectrum for structure X_2, namely the chemical shift for vertex j (−CH═ in the aliphatic ring, see Figure 3). The first environment level of C_j



is found in CT; however, the second level



is not. Thus, for C_j the CT provides only the first level information: there were 8566 vertices with such an environment in the database, their δ range from 62 to 165.5 with the mean value of 120 ppm. Taking into account the real $\delta = 104.7$ one should consider the result as unsatisfactory. And still the database contains structure R_4 which may help to make a much better prediction:



R_4

Compound R_4 is the structural analog of X_2 *respective* vertex *j*. It has been found with an aid of a *classifier*.[23] The classifier

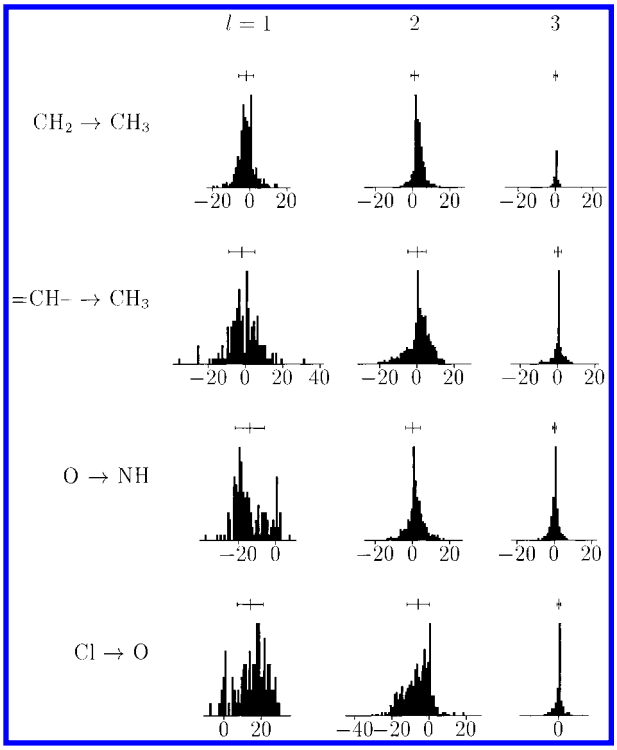


Figure 4. Chemical shift influence histograms for some substitution kinds (*l* denotes the environment level).

allows directly addressing structures with similar (not obligatory coinciding) environments of any vertex of interest. Indeed, the second environment levels of C_j and the corresponding carbon in R_4 differ only by the substitution of NH by ether O. Both functions, however, have a similar influence on the resonating carbon atom δ, so the value of 104.1 found for R_4 could be a good approximation. So if the direct match is absent from CT, it may be useful to look for a similar environment there.

Let us generalize this technique for arbitrary vertex types. To find all relevant substitutions, we have enumerated all pairs of environments differing only by a single vertex type at the last level. Each pair represented a particular case of substitution, characterized by the exchanged vertex types pair, the level number *l*, and the beheld δ difference (d_δ). The exhaustive enumeration of all such pairs allowed us to collect d_δ distributions for all kinds of substitution taking place in CT.

Let us adduce some characteristic distribution histograms. Indeed, Figure 4 attests an omissible consequence of the substitution NH → O at the second environment level. However, at the first level the distribution shape points to a significant spread effect. A quite predictable and stable effect is shown by the substitution $CH_2 \rightarrow CH_3$, whereas ═CH− → CH_3 exhibits an opposite feature that may be expected in advance due to different hybridization of exchanged vertices. It is worth noting that the substitution effects usually have reverse signs at the first and second levels (it is especially evident in example Cl → O). Finally, all the distributions show the minimal effect of any single substitution at the third level.

The d_δ distributions are used in a shortened form of the *allowed substitution table*, AST. For every substitution kind AST has only two parameters of a d_δ distribution: the mean value $\bar{d}_\delta$ and the mean deviation $\tilde{d}_\delta$. The AST involves far

Table 1. Part of AST for Carbon Vertices on the First Two Environment Levels

	$l = 1$		$l = 2$	
	$\bar{d}_\delta$	$\tilde{d}_\delta$	$\bar{d}_\delta$	$\tilde{d}_\delta$
$CH_2 \rightarrow CH_3$	−2	4	1	2
$CH \rightarrow CH_3$	−4	5	2	3
$C \rightarrow CH_3$	−8	6	3	3
$CH \rightarrow CH_2$	−2	4	1	2
$C \rightarrow CH_2$	−5	5	2	3
$C \rightarrow CH$	−4	5	1	2
$-CH= \rightarrow CH_2=$	3	5	1	2
$>C= \rightarrow CH_2=$	11	12	3	3
$>C= \rightarrow CH=$	1	5	0	3

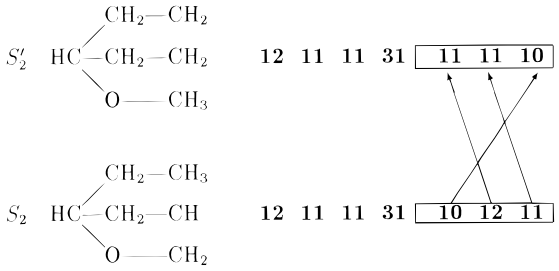


Figure 5. Vertices comparison in matching environments.

from all possible substitutions (all omitted ones are forbidden). On preparing AST all vertex types were broken into five groups for chemical reasons: aliphatic carbon vertices (CH_3, CH_2, CH, C); carbon vertices in sp^2 hybridization ($CH_2=$, $-CH=$, $>C=$); heteroatoms without multiple bonds (F, Cl, Br, J, OH, O, NH_2, NH, N, SH, S); nitrogen, oxygen, sulfur with a double bond (O=, HN=, −N=, S=); and all the rest of the vertices. The separation aims banning substitutions between vertices from different groups. Unfortunately, the reverse feature (the ability to make any changes within a group) is not always met, since some substitutions (e.g., Cl → O at all levels) were rejected because of too spread out or low populated distribution (the authors made these subjective conclusions from the histogram appearance). The closure of these gaps, however, is a technical problem that can be solved by analyzing a larger and more representative CT. As an example, in Table 1 a complete AST part related to carbon vertices is adduced.

The AST is used if the second or first environment of the vertex of interest (further, the *target* environment) is not found in CT. Then all records for these level environments with only the last level differing from the target environment are scanned through. The ISAM implementation of CT guarantees all such records to be within some local directly accessible region, so the scanning can be fast and efficient. Each record in this region is considered further if its environment *matches* the target one, i.e., one can be transformed into another only by means of allowed substitutions.

Let the target environment S_2 (Figure 5) be absent from CT that entails sequential reading of CT aiming matching environments. Let the current record answer query environment S_2' which differs from S_2 only at the last (second) level. The algorithm compares two environment codes, i.e., record keys, taking into account only terminal sections related to the last level (coincidence of the starting sections is the binding condition whose violation causes the scanning to stop). In our case only three last symbols (**10 12 11** for the target and **11 11 10** for the query environment) are considered. Here these symbols denote just vertex types (in a common case there can be ring closures too, see ref 23 for details). The goal is to find for each symbol in the first section a corresponding symbol in the second one.

The first symbol **10** (type of vertex CH_3) exists in both sections, though in different places. This difference is accounted by means of *penalty p*—some measure of environments dissimilarity. By default the place difference penalty equals 2, so after the first step $p = 2$. Then goes **12** (CH). The second section has not the same symbol, though there is **11** (CH_2) which can serve as a change for **12** (see Table 1, row CH → CH_2). Assured with no exact match, the algorithm stops at the first allowed change found—in our case it is the first symbol **11**. So the pair **12** → **11** is charged by the place penalty too. Additionally, two values $\bar{d}_\delta = +1$ and $\tilde{d}_\delta = 2$ are acquired from AST. Mean deviation $\tilde{d}_\delta$ is added to the total penalty, and offset d_δ composes another value known as a *sum correction s* of a query environment. Finally, for the last symbol **11** there remains single **11** on the second place that means two more penalty points. So environment S_2' is accepted to match S_2 with total penalty $p = 8$ and sum correction $s = +1$ ppm.

There can be (and usually are) many matching fragments, each with its own *p* and *s*. Whenever the sum correction means an expected difference between δ's of the target and matching fragments, the penalty serves as a discredit measure of such an expectation. Finally, the chemical shift of an environment outside CT is calculated by the formula

$$\delta = \frac{\sum (\delta_i + s_i)/p_i}{\sum 1/p_i}$$

Let us call the described prediction method as an *adaptive* one. Its sense boils down to the use of analogous (matching) environments, when the target one is absent. This method is evidently more complex and slow than the direct use of CT. However, this disadvantage can be tempered by writing adaptive prediction results into a specific ISAM table which is searched each time before starting the extensive scanning of matching environments in the main CT. As the prediction is usually held for somehow similar structures, the use of an additional table turns out to be very efficient. With data accumulation in this table the extra time expenditure due to the adaptive character of the prediction becomes almost insensible.

6. RESULTS AND CONCLUSION

So the present work suggests a general scheme of structure elucidation based on a combined use of spectral databases and ^{13}C NMR correlation tables (Figure 6). Here databases serve as the main information sources, while CT performs mainly screening functions. As spectral databases contain a larger amount of data than any sources of processed information (e.g., CT), one can expect definite advantages of our method over those with no provision for direct use of raw specral data.

At the same time contents and quality of databases are not responsible for the method's efficiency in a full measure. Particular algorithms play a complementary role too. Some of their features, not reflected in the scheme, should shortly

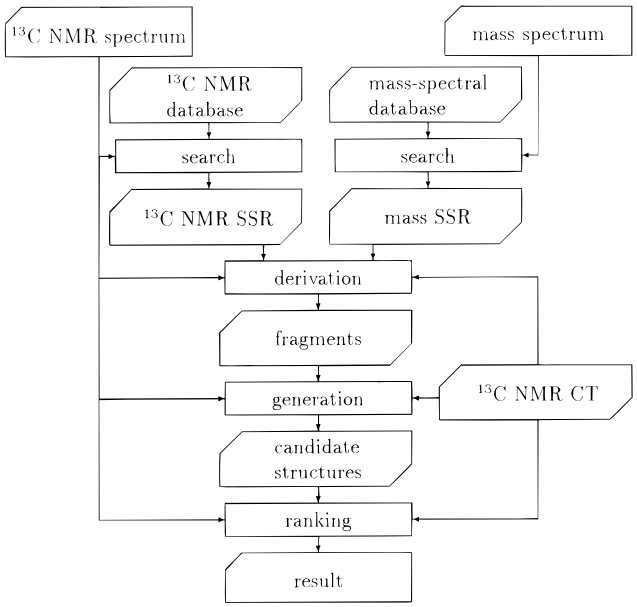


Figure 6. The general scheme of structure elucidation based on combined use of spectral databases and ^{13}C NMR correlation tables.

be repeated here: fitting of analyzing structures to spectrum *X*, fragments combination, and spectral bans on bond formation. Adaptive prediction is important almost at every step, besides, it can be used on its own.

Adaptive prediction efficiency can be illustrated by structure X_2 (Figure 3). Ten carbon vertices in X_2 have no second or even first level environments described in the used CT. Data on these vertices compose the following table, where column δ_x is related to experimental values in ppm, δ_r—found in CT for *l*th environment level, δ'_r—results of the adaptive method use

vertex	δ_x	δ_r	l	δ'_r
a	191.2	196.2	1	192.1
b	104.1	127.9	1	109.9
c	157.6	141.7	0	155.1
f	95.3	122.8	1	109.9
g	164.5	145.3	1	150.9
h	90.1	86.3	1	88.7
i	170.3	151.0	1	154.5
j	104.7	120.0	1	105.9
l	39.5	39.4	1	44.1
m	35.4	38.4	1	38.4

The method is evidently not free from faults: vertex *m* has not improved its results, and vertex *l* has even slightly spoiled them. However adaptive prediction as a whole turns out to be quite useful—without it structure X_2 would drop from second place to the middle of the ranked list of candidate structures.

As additional examples Figure 7 shows the structures of 10 unknown compounds determined by the proposed computer method using a molecular formula, mass, and ^{13}C NMR spectra. For every task as a characteristic of a solution outcome in Figure 7 we give the following data: the position of an unknown structure in the ranked list of generated structures, the total number of generated structures, and the time elapsed on PC 486-DX2 66 MHz. Though in many tasks a correct structure does not occupy the first place, one can hardly expect better results taking into account the available spectral data. A detailed examination of such cases shows

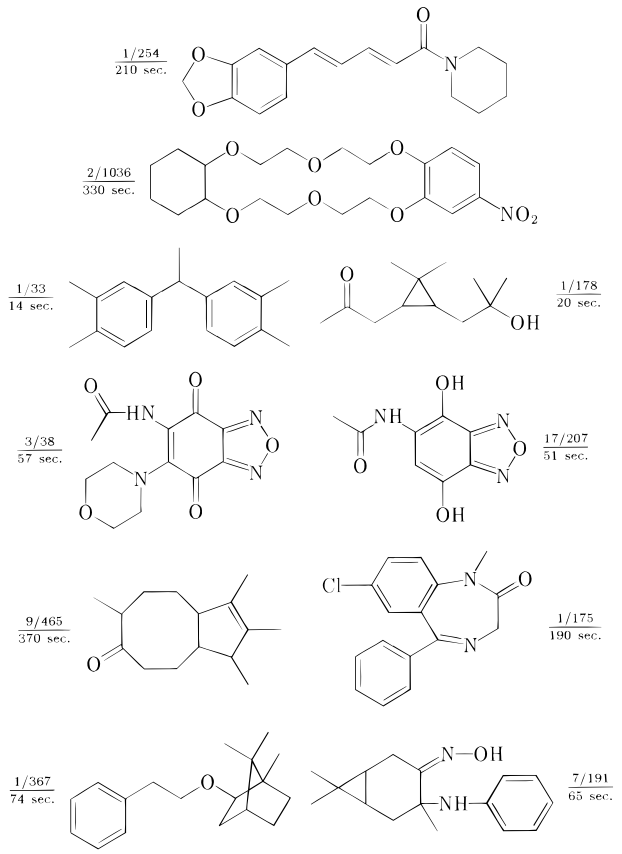


Figure 7. Some structural problems successfully resolved with the aid of ChemArt system using mass and ^{13}C NMR specrta and MF of unknown compounds. The notation used: $(n_x/n)/t$, where n_x is the place of an unknown compound in the ranked list of candidate structures, *n* is the list size, *t* is time elapsed for a task on PC 486-DX2 66 Mgz.

that the use of other spectra (^{1}H NMR and IR) would have a significant screening effect on both the fragments derivation and ranking stages. This remark especially relates to presented heterocyclic compounds, for which mass and ^{13}C NMR spectra are evidently insufficient for an unambigous solution.

The total number of generated structures and elapsed time are less representative as they depend on a search result contents and numerous user-defined parameters which may vary from task to task. In this work we use search procedures which take into account distinct spectral features (chemical shifts and multiplicities for ^{13}C NMR spectra, *m/z* of ions, masses of neutral losses, and peaks intensities for mass spectra), tolerances of the features coincidence, and allow a user to apply different search methods (forward, reverce, combined) and match factors.[12,27] In any case a spectral search result is a list of selected compounds ranked by a spectral match factor. In our experiments up to 20 first structures from each (^{13}C NMR and mass) search result ($\leq$40 altogether) were used for the subsequent analysis.

In our case structural problems handling was always held within the shell of the highly modular ChemArt system,[24] which provides both the ergonomic user interface and convenient way of algorithms introduction and adaptation. It should be noted however that most problems in Figure 7 were not solved fully automatically and required the human interference both on the stage of parameters choice and on the making decision as to whether the structures, proposed

by the system, contain the correct one (if the right solution was not known in advance).

Dealing with numerous problems convince us of good potentiality of the proposed approach. Its efficiency is especially evident when the spectral search succeeds in finding close structural analogs of an unknown compound. The opposite cases, however, often call for some iterations to pick up appropriate search and retrieval parameters. Our plans include elimination of this shortcoming through a wider and more intelligent automatic parameters adjustment. Extensive development of the method by incorporation of IR and ^{1}H NMR information is also perceived.

ACKNOWLEDGMENT

This work was supported by the Russian Basic Research Foundation (Grant 97-03-33514). The authors thank Dr. L. A. Ostashevskaja (Novosibirsk State University) and Prof. K. Funatsu (Toyohashi University of Technology) for the provision of interesting structural problems.

REFERENCES AND NOTES

(1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986.
(2) Zupan, J. *Computer-Supported Spectroscopic Databases*; Hastled: New York, 1986.
(3) Gray, N. A. B. Artificial Intelligence in Chemistry. *Anal. Chim. Acta* **1988**, *210*, 9−32.
(4) Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Crandell, C. W. The DENDRAL Project: Resent Advances in Computer-Assisted Structure Elucidation. *Anal. Chim. Acta* **1981**, *133*, 471−497.
(5) Funatsu, K.; Sasaki, S. Resent advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Function for Examination of Candidates. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.
(6) Elyashberg, M. E.; Martirosian, R. R.; Karasev, Y. Z.; Thiele, H.; Somberg, H. X-PERT: a user-friendly expert system for molecular structure elucidation by spectral methods. *Anal. Chim. Acta* **1997**, *337*, 265−286.
(7) Shelley, C. A.; Munk, M. E. CASE, a computer model of the structure eluciation process. *Anal. Chim. Acta* **1981**, *133*, 507−516.
(8) Schaller, R. B.; Munk, M. E.; Pretsch, E. Spectra Estimation for Computer-Aided Structure Determination. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 239−243.
(9) Warr, W. A. Spectral Databases Chemom. *Intell. Lab. Syst.* **1991**, *10*, 279−292.
(10) Bremser, W.; Grzonka, M. SpecInfo−A Multidimensional Spectroscopic Interpretation System. *Mikrochim. Acta* **1991**, *2*, 483−491.
(11) Neuderd, R.; Penk, M. Enhanced Structure Eluciation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 244−248.
(12) Lebedev, K. S.; Derendyaev, B. G. Computer Methods for the Solution of Structural-Analitycal Problems with the Help of Data Bases on Molecular Spectroscopy (MS, IR, NMR). *Chem. Sustainable Dev.* **1995**, *3*, 249−265.
(13) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Eluciation−A Spectroscopist's Dream Comes True. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.
(14) Gatilova, V. P.; Korchagina, D. V.; Gatilov, Yu. V.; Bagryanskaya, I. Yu.; Barhash, V. A. Molecular rearrangements of α-humulene and its 6,7-epoxide in superacids (in Russian). *Zhurnal Organicheskoi Himii* **1991**, *27*, 2301−2318.
(15) Haraki, K. S.; Venkataraghavan, R.; McLafferty, F. W. Prediction of Substructures from Unknown Mass Spectra by the Self-Training Interpretive and Retrieval System. *Anal. Chem.* **1981**, *53*, 386−392.
(16) Lebedev, K. S. Use of IR and Mass Spectroscopic Databases to Establish the Structure of Organic Compounds. *J. Anal. Chem. Engl. Tr.* **1993**, *48*, 603−611.
(17) Derendjaev, B. G.; Nekhoroshev, S. A.; Lebedev, K. S.; Kyrshansky, S. P. Computer-Aided Molecular Formula Determination from Mass, ^{1}H, and ^{13}C NMR spectra. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 255−260.
(18) Munk, M. E. *The Role of NMR Spectra in Computer-Enhanced Structure Elucidation. Computer-Enhanced Analytical Spectroscopy*; New York, 1992; pp 127−149.
(19) Seger, C.; Jandl, B.; Brader, G.; Robien, W.; Hofer, O.; Greger, H. Case studies of CSEARCH supported structure eluciation strategies: lupeol and a new germacrane derivative. *Fresenius J. Anal. Chem.* **1997**, *359*, 42−45.
(20) Dubois, J. E.; Carabédian, M.; Dagane, I. Computer-Aided Elucidation of Structures by Carbon-13 Nuclear Magnetic Resonance. *Anal. Chim. Acta* **1984**, *158*, 217−233.
(21) Molodtsov, S. G. Generation of Molecular Graphs with a Given Set of Nonoverlapping Fragments. *Math. Chem. (MATCH)* **1994**, *30*, 203−212.
(22) Molodtsov, G. G. Computer-Aided Generation of Molecular Graphs. *Math. Chem. (MATCH)* **1994**, *30*, 213−224.
(23) Strokov, I. A. Compact Code for Chemical Structure Storage and Retrieval. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 939−944.
(24) Strokov, I. I.; Lebedev, K. S. A New Modular Architecture for Computer Systems in Chemistry. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 741−745.
(25) The source code and documentation on JoKey file handler is freely available at ftp://ch-inf.nioch.nsc.ru/strokov/jokey/.
(26) Bremser, W. HOSE−A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(27) Lebedev, K. S., Cabrol-Bass, D. New Computer-Aided Methods for Revealing Structural Features of Unknown Compounds Using Low Resolution Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 410−419.

CI980184P