Problem-Solving Methods in Computer-Aided Organic Structure Determination

ZDZISŁAW HIPPE

Department of Physical and Computer Chemistry, Technical University, 35-041 Rzeszów, Poland

Received January 28, 1985

Common features of various identification algorithms for organic structure/substructure determination have been briefly discussed. Also, some comments on existing programming environment for computer-assisted structure determination are given. Then, a novel approach to problem solving in elucidation of organic structures, the Independent Identification by Artificial Intelligence, IIAI, has been briefly outlined.

INTRODUCTION

Elucidation of organic structures is an important step in the investigation of natural substances and in research on syntheses of new compounds: apart from its cognitive significance, very often it is the controlling factor of the overall direction of the experiment itself. Recently, structural identification is achieved by empirical interpretation of data from fracture of excited molecules (mass spectrum, MS) or by interpretation of molecular spectra like nuclear magnetic resonance (carbon-13 or proton, ^{13}C NMR or ^{1}H NMR, respectively), infrared (IR), Raman (RA), or ultraviolet (UV).[50] The process we are talking about may be regarded in the sense of identification of the detailed structural formula or in the sense of recognition of its principal fragments (often called substructures) to define—for instance—a chemical class of which the investigated substance is a member or to make deductions about major structural changes, which have occurred during a given reaction or process. Within the last 10–12 years, various examples of the computer-assisted interpretations of molecular spectra were pointed out. From simple algorithms (and programs), enabling identification of a very limited set of compounds,[1] we arrived at large, heuristic computer program systems, performing extremely complicated research tasks during estimation of the exact structures of unknown chemical compounds of complex constitution. Those systems may be regarded as "expert systems", which—according to recent definition[2]—under appropriate circumstances capture the skill of the expert in a given domain, in the form of computer software.

As in any other newly created scientific discipline, also in computer-assisted structure determination we may observe some divergent definitions of notions and theoretical tools, used by various authors. For instance, it is suggested[3] that structural identification, aided by the computer, consists of the application of pattern recognition methods, of the search through a library of standard reference spectra, or of the application of artificial intelligence programs, for modeling of human intellectual processes during the interpretation of a given spectrum. But, accordingly to the most recent classification accepted by many authors, all methods mentioned above belong to the area of artificial intelligence (AI).[4] It comes from the fact that even the manual search through a collection of standard spectra is an example of typical intelligent behavior of a human being, especially when the search process is self-guided by deductive feedback. In the present article, we leave behind the questions of pattern recognition, but we would rather stay with the problem-solving methods (as part of AI) in the process of empirical interpretation of an unknown spectrum (any type) by simulating the action of a human analyst. It should be emphasized that opinions and conclusions presented here do not serve for comparison of the effectiveness of existing program systems for computer-aided structure determination[5-21] but only for the general progress of the problem-solving method in the area stated.

Prof. Zdzisław Hippe graduated from the Technical University of Łódź and from the Technical University of Gdańsk in chemistry. He received his Ph.D. in technical sciences in 1965 from the Technical University of Gliwice and his habilitation (also in technical sciences) in 1968, again from the Technical University of Łódź. In 1969 he joined the Technical University of Rzeszów, where he is currently the Head of the Department of Physical and Computer Chemistry. He teaches computer science for chemistry students. Since 1975, he is a professor of technical sciences; he has lectured many times abroad. His main research activities include artificial intelligence in chemistry, particularly in the design of organic syntheses and in structure elucidation. Recently, he is serving as a member of the Committee of Analytical Chemistry and the Committee of Computer Science of the Polish Academy of Sciences.

GENERAL SOLUTIONS IN COMPUTER-AIDED STRUCTURE DETERMINATION

The thorough overview of accessible literature suggests that almost all existing algorithms (and programs) for computer-assisted determination of organic structures/substructures solve the problems by copying the scientist's brain process, which is going on during the interpretation of an unknown spectrum.[51] We should bear in mind that this intellectual process depends to some extent on the spectral technique considered. On the other hand, the overall efficiency of such algorithms (for interpretation of any type of spectrum) is strongly dependent on the experience of the algorithm designer. In other words, each problem-solving algorithm for structure determination embodies and reflects the knowledge and intuition of the spectroscopist. But immediately, we touch the main difficulty involved: how to translate the chain of sophisticated intellectual subprocesses into a rigid sequence of computer operations, formulated in a way suitable for machine processing? Inspection of the Table I shows clearly the weak points of

0095-2338/85/1625-0344$01.50/0 © 1985 American Chemical Society

Table I. Comparison of Capabilities of Man and Computer in Structure Determination

	man	computer
memory	slow, unreliable	fast, reliable
evaluation	global	local
search of space of solutions	heuristic	exhaustive
biasing of deductive process by experiment	yes	no

Table II. Main Features of Computer Program Systems for Structure Determination (Based on *One* Spectral Technique and on Arbitrary Combination of Spectral Methods)

feature	type of system	
	monomethod	combined
requested design of experiment	expertized	less rigorous
interpretation of data	profound	superficial
data set	small	large
programming cost	low	high

computer-aided structure determination. In the current status of problem-solving methods in AI, the change of evaluation area and the implementation of deductive biasing by the experiment seem to be distant goals. Besides, the space of possible solutions for structure determination is far too large to search exhaustively, except for trivial problems. Thus, we need to develop special knowledge (theories and facts) for heuristic reduction of the solution space.

At the present time, two distinctly different approaches to problem solving in structure determination are mentioned. The first relies on the interpretation of one spectrum only (of a given type, say MS, ^{13}C NMR, ^{1}H NMR, IR, etc.); we may call this case monomethod systems. The second approach is based on the interpretation of a set of spectra of the unknown substance, but taken with various spectral methods. This procedure, if realized by the computer, may be regarded as a base for so-called combined systems for structure determination (also known as integrated[22] systems), designed with special philosophy, offering many sophisticated possibilities for users. Before we enter the detailed discussion of the internal structure of integrated systems (see Integrated Systems For Structure Determination: Their Advantages and Disadvantages), let us compare their characteristic features with those of monomethod systems (Table II). Briefly speaking, whereas one may employ a carefully selected spectral method in investigating a given problem, interpreting data in considerable depth, the integrated systems (almost any combination of spectral techniques is used, according to available instrumentation) do processing of experimental data rather superficially. However, the final results of structure determination seem to be more informative and reliable, because this procedure uses the fact that various spectroscopic methods give structurally significant pieces of information, which may mutually confirm, complete, and/or eliminate. The improvement of results in structure elucidation while combining the different spectral techniques was announced already in 1969 by Isenhour and co-workers.[23]

In the following part of the article, basic types of algorithms used in monomethod systems will be briefly discussed. The same types of identification algorithms have been applied in integrated systems to perform the first step in all five during determination of exact structure.

IDENTIFICATION ALGORITHMS

The identification algorithms serve for automatic (by the computer) interpretation of a given spectrum, in order to recognize the detailed structure of the compound being analyzed or to detect as precisely as possible the structural fragments (substructures) that go to form the molecule. The relation F is here postulated:[23]

$$\text{spectrum} = \sum\text{partial spectra} \xrightarrow{F} \text{structure} = \sum\text{partial structures}$$

which tends to show the possibility of receiving the prediction on full structure or on partial structures (substructures) only, according to entropy of information on the investigated molecule.

For the sake of simplicity, we assume that the problem-solving algorithms for structure determination (and hence for interpretation of a spectrum) are not dependent on the spectral technique used. Thus, the discussed algorithms may be classified into three main groups: (1) library-search algorithms, (2) network algorithms, and (3) matrix algorithms.

Library-search algorithms correlate with the comparison (done on numerical basis) of an unknown spectrum with those in standard collection stored on a machine carrier of information, say Winchester or floppy disks. Hence, the standard reference spectra are converted in some way to digital representation suitable for computer processing. They may be stored as vectors (each element stands for band/peak location or for intensity of subsequent bands, provided the step of digitalization is constant) or as 2D matrices. The search process runs in three distinct steps: (a) conversion of an unknown spectrum into the same digital form used throughout the data base, (b) comparison of numerical representation of an unknown spectrum with numerical representation of subsequent reference spectra, and (c) output of a list (preferably not extensive) of spectra identical with and/or similar to the unknown, according to an a priori fixed criteria of similarity. The list of spectra should be sorted out with decreasing similarity and each spectrum provided with a so-called fitting factor, usually adjusted to 1.000 for exact likeness of the spectra compared (unknown and reference). A final decision in the choice of a proper spectrum (proper structure) is made by a chemist. Characteristic features of all search algorithms are very simple programming and proportionality of machine time for unitary search to the size of a data base. There are many different solutions of the organization of the search process itself, in the sense of computer science. Usually, the sequential search (i.e., comparison of an unknown spectrum with each subsequent spectrum in standard reference collection) is avoided. The machine time may be cut down efficiently by usage of inverted data files (what resembles the search through general index and then through the text) or doubly inverted files[24] (index → index → text). The search systems for structure determination described in the literature apply a so-called positive search (it consists of selecting those parameters of an unknown spectrum that have to be present in a standard spectrum searched from the data base) and also a negative search[25] (here those parameters should be specified that are not present in the spectrum retrieved).[52] A combined approach (negative + positive search) has also been reported.[26] Apart from these techniques for searching files of spectral data, the comparison of an unknown and reference spectra may be done in the forward[27-28] or in the reverse sense.[29] In a reverse search the question is asked "is the unknown substance a particular compound X?", and therefore, the library spectrum of the compound X is used as the basis for comparison with the unknown one. In the forward search the question is whether reliable determination of the unknown structure may be possible, even under the condition that the interpreted spectrum is severely distorted, e.g., by background absorption or unresolved components (in a band complex). Therefore, in all cases where we cannot expect a pure unknown spectrum (for instance, when mixtures are being investigated), a reverse search method may be advisable. A combined forward–reverse search developed for the data base of mass spectra[30] was

reported to give satisfactory results.

The library-search algorithms supply two types of results. When the unknown spectrum has the exact match within spectra stored in the data base, we get exact determination of the unknown structure (i.e., the structure corresponding with the standard reference spectrum retrieved may be assigned to the unknown compound, with high degree of probability, say 0.9999). Alternatively, when the unknown spectrum does not have the exact match in the collection of standard reference spectra, we may obtain information on the chemical class of the molecule being analyzed and/or about its substructures. The correct information on substructures contained in the investigated molecule may be readily retrieved if the content of standard reference spectra collection reflects the user's interest. Therefore, deserved popularity gets small data bases supplied even on floppy disks (with various types of spectra), covering the most important domains of industrial chemistry, for instance, drugs, polymers, biocides, etc. However, we should keep in mind that the performance of data-base systems depends not only on the library-search algorithms used but also—to some extent—on the user's experience.[31]

As was previously mentioned, the problem solving of structure determination by library-search algorithms consists in modeling the behavior of a human analyst, who interprets an unknown spectrum by means of the manual with reference spectra. Accomplishment of such a formally simple task as the comparison of two objects (in this case an unknown spectrum and a reference spectrum taken from the manual) requires, however, some experience, knowledge, and intelligence! These factors have essential influence on performance of the library search and on quality of results received. Also, various heuristics or even simple ad hoc programming shortcuts or complex logical operations have been applied to improve the effectiveness of the library-search algorithm and to reduce the solution space, thus decreasing the machine time needed for identification of structures/substructures. But it should be emphasized once again that *all* these programming tips and hacks are also copying the behavior of a human expert. Many authors look for such problem solving almost intuitively, possibly not even recognizing that they follow these lines. For instance, Zupan[32] announced a library-search system (and algorithm) in which the user may enter the most intense band parameters (from an unknown spectrum) to generate a spectra subset of decreased size. Then, entering another band parameters (location and intensity), a subsequent set of spectra is anew generated (the size of the second subset is again decreased); all of them contain both selected bands, etc. In reality, the algorithm described copies exactly the behavior of the analyst, who—while interpreting an unknown spectrum—takes into account successive bands of highest intensity.[53] Similarly, the so-called interpretive search algorithm[34] (which not only identifies a carbon-13 NMR spectrum that is identical with or similar to a spectrum of an unknown compound but serves to locate reference structures in the library that contain one or more identifiable substructures in the data base that model some substructures of the unknown) displays exact analogy to the brain process of a very experienced spectroscopist, who is going through the reference spectra manual and simultaneously considering the cross-correlations with feasible substructures for the substance being tested. We see again the quality of the library-search algorithm (and other types of algorithms too!) depends mainly on the knowledge and experience of the algorithm designer and reflects his/her creative ability.

Taking into account that the largest data-base collections of standard reference spectra amount to less than 0.8% of known organic substances,[35] the aim of finding more general solutions than simple library search is obvious. This is accomplished by other types of identification algorithms.

Network or AND/OR tree algorithms reflect the human reasoning during the interpretation of an unknown spectrum with the usage of correlation tables. Evidently, almost any spectroscopist has his/her own way (algorithm) for interpretation of an unknown spectrum. But, some useful rules or even algorithms expressed in words have been described, for instance, in references 36 and 37. According to these general rules, this type of the interpretation process relies in any case on raising binary questions (only one of two possible answers, "YES" or "NO", is allowed). The questions are usually connected with checking whether a particular band (peak, multiplet) does exist within a specified spectral region. Also, negative binary questions may be put forth if necessary, for instance, whether a given spectral region is empty (no band contained), etc. In addition to the binary questionning, the network algorithms display another common feature. Namely, in any of them the unceasing "jumping" from one spectral region to another may be observed. To explain this statement, let us consider the interpretation of an IR spectrum. The first question may be "does the spectrum contain a band of medium intensity, in the region 3150–3000 cm^{-1}?" If the answer is YES, the next question jumps toward the spectral region of out-of-plane deformation vibrations ν_{C-H}, just to confirm the former finding (aromatic and/or unsaturated substructures). And the next jump may be toward the region of aromatic substitution overtones, etc. We see that this jumping, speaking in a different way, indicates the sequence of binary questions in the identification algorithm. Therefore, the algorithms discussed represent a network of binary trees, with very complicated logical interconnections using Boolean algebra.

The application of tree-like algorithms for computer-aided structure determination comes from information theory. This theory also helps in the proper choice of binary questions and their sequence. Suppose we are given a set of alternatives (in our case, substructures) that may be detected by a network algorithm:[54]

$$X = \{x_1, x_2, x_3, ..., x_N\}$$

The binary questioning for specified spectral features of the alternatives $x_1, ..., x_N$ enables one to guess which alternatives (substructures) from all allowed are expected. The same result may be achieved, but probably in a more optimal way, if the binary questions are within the limits of spectral bands (multiplets, etc.) really existing in the spectrum analyzed. A fixed manner to place questions in a suitable sequence is called the identification system of elements in the set X. In structure determination we meet sets in which alternatives (substructures or spectral features) do not have the same probabilities. Thus, for each element in the set X, the respective probabilities of alternatives x_i may be determined:

$$P\{x_i\} = p_i \qquad i = 1, 2, 3, ..., N$$

The inclusion of the probability concept in structural identification leads to a very important expression for evaluating the efficiency of a system. The number of questions $E(S_k)$ in an identification system S_k, for recognition of elements in the set X, is given by

$$E(S_k) = \sum_{i=1}^{N} S_k(x_i) p_i$$

as the criterion of optimization. Clearly, the system S_1 is better than S_2, if

$$E(S_1) < E(S_2)$$

For proper choice of questions and their sequence, we may apply some general rules, for instance, those of Huffman[38] and Kraft.[39]

Table III. Results of Static Interpretation of IR Spectrum of 2-Aminoethanol

no.	substructure detected	its characteristic spectral features	identification factor (%)
1	ALKYL-NH2	3360.7, 3280.8, 1590.33, 1076.16, 860.25	95
2	ALKYL-OH	3360.7, 1390.60, 1038.12, 950.20	95
3	ALKYL-	2920.5, 1450.45	94
4	-CH2-	2860.5	90
5	ALKYL-NH-	3280.8, 1590.33, 1163.73, 1076.16, 860.25	88
6	>S=O	1038.12	88
7	ALKYL-O-	1076.16	65
8	>C=N-OH	3280.8, 950.20	52

Characteristic attributes of computer program systems based on network algorithms are a very short machine time for one search and large programming efforts that rise exponentially with the size of the set of identified substructures. But an entirely new possibility for overcoming of this drawback is the method of self-programming (the program is written by the computer itself) of binary trees.[40] The process requires some human guidance, namely, only a small portion of the program (closing part which consisted of instructions for printing) should be written by a programmer. The self-programming of the computer for structure determination was applied for the first time in the designing of identification system POLYMER.[41] Research on automatic program-writing procedures was later developed with the use of new logic,[42] in which the logical state of truth is labeled by 1 whereas falsehood is denoted by 2. In this way a family of keys generated for the tree of specified order[55] may be interpreted as line numbers in such codes as FORTRAN or BASIC.

Matrix algorithms, introduced a short time ago for the interpretation of IR data,[43,44] have their internal structure formally similar to tree-like algorithms.[56] However, the main difference is in the sequence of binary questions: all bands are assigned sequentially, beginning from the highest wavenumber. The information about substructures detected is stored in the substructure matrix (for the sake of calculation speed, this is in fact a vector). Each element of the matrix stands for a respective substructure (detected by the algorithm), whereas the numerical value of the element stands for the probability of the substructure inclusion in the investigated compound. At the beginning of the interpretation process, all elements of the vector are zeroed. Then, the status of the characteristic spectral parameters of each consecutive band (peak) is checked against a carefully designed adjustable binary decision function,[57] of the general form:

$$G(J) = G(J)_{\text{init}} + A + B + C + D$$

where $G(J)$ is the value of the Jth element of the substructure vector, calculated by means of the decision function, $G(J)_{\text{init}}$ is initial value of the Jth element, when only some logical conditions are met (except those specified in the terms A, B, C, and D), and A, B, C, and D are additional terms, describing the goodness of the fit between the spectral parameters of the assigned band and the "ideal" band for location (A), intensity (B), diagnostic power (C), and occurrence probability of other bands in the region tested (D), respectively.

3360.7	1590.33	1230.77	950.20
3280.8	1450.45	1163.72	860.25
2920.5	1390.60	1076.16	620.70
2860.5	1353.45	1038.12	510.73

Figure 1. IR vector of 2-aminoethanol. The number 3360.7 means the location of the IR absorption band is 3360 cm^{-1} and the transmittance equals 7%.

We see that the decision function—although binary from the assumption—has the special property of being adjustable. The final result is not only the alternative a given substructure is present (or absent), but in fact, we get the discrete value for $G(J)$, which ranges from 0 to about 1111, according to quality of fitness between the real band and the ideal one. The $G(J)$ value is assumed to be proportional to the inclusion probability of a given substructure within the investigated molecule. After the last band (peak) is interpreted, the absolute values of all elements of the substructure vector are sorted and printed, as a list of substructures detected, sequenced according to decreasing probability of identification. Substructures with $G(J)$ equal to zero are of course not printed and treated as undetected.

The method of saving the information during the interpretation of a spectrum in the substructure vector enables very convenient manipulation of the $G(J)$ values for any substructure. For instance, knowing from the history of the sample that it does not contain an aromatic ring, the respective $G(J)$ element may be interactively zeroed very easily, hence scratching out the substructure from the list. Inversely, the value of a given $G(J)$ element may be simply rised, provided we have some information about substructure (or heteroatoms) related to it. Thus, the personal intuition and knowledge of the user about the investigated substance may interactively affect the results of structure determination.

The most recent development in the field of the matrix algorithms consists in the application of the so-called static/dynamic method of interpretation. Though data published till now were exclusively devoted to the computer-aided interpretation of IR spectra, this combined method may readily be applied to other techniques.

The course of action of the algorithm is explained by means of the simple example[47] of the IR spectrum of 2-aminoethanol. In the first step (called static interpretation), the IR vector (see Figure 1) was dealt with the matrix algorithm; the results are gathered in the Table III. In the third column we have the information (extracted from the computer program); which bands, as logical conditions, were taken into account for the preliminary assignment of particular substructures. Some of them were used twice or even more times. In the second step (dynamic), the set of substructures detected previously is analyzed in a specially established sequence. The intermediate results are showed in Table IV, where also the concept of blocking and deblocking of bands, while interpreting the spectrum, is presented. Namely, after analysis of the sub-

Table IV. Course of Dynamic Interpretation of 2-Aminoethanol Spectrum

no.	substructure tested	bands blocked	band deblocked	not blocked bands in IR vector
1	ALKYL	2920, 1450		3360, 3280, 2860, 1590, 1390, 1353, 1230, 1163, 1076, 1038, 950, 860, 620, 510
2	-CH2-	2860		3360, 3280, 1590, 1390, 1353, 1230, 1163, 1076, 1038, 950, 860, 620, 510
3	ALKYL-NH2	3360, 3280, 1590, 1076, 860		1390, 1353, 1230, 1163, 1038, 950, 620, 510
4	>C=N-OH			1390, 1353, 1230, 1163, 1038, 950, 620, 510
5	ALKYL-NH-			1390, 1353, 1230, 1163, 1038, 950, 620, 510
6	ALKYL-OH	1390, 1038, 950	3360	1353, 1230, 1163, 620, 510
7	ALKYL-O-			1353, 1230, 1163, 620, 510
8	>S=O			1353, 1230, 1163, 620, 510

Table V. Final Results of Static/Dynamic Interpretation of 2-Aminoethanol Spectrum

no.	substructure detected	identification factor (%)
1	ALKYL-NH2	96
2	ALKYL	94
3	CH2	91
4	ALKYL-OH	85

structure ALKYL (it may be even a short chain), two bands are blocked (reserved). Similarly, the substructure –CH2– eliminates the 2860-cm^{-1} band from the IR vector; ALKYL-NH2 eliminates another four, etc. The next two substructures cannot be detected, as the bands related to them have been already used. Somewhat different situation is with the substructure ALKYL-OH. Here, from four bands required for its recognition (see Table III), three (1390, 1038, and 950 cm^{-1}) are at the moment not used up. Also, the identification factor (IF) from the static interpretation is very high. For these reasons, the "missing" band 3360 cm^{-1}, already reserved for ALKYL-NH2, is deblocked, i.e., reassigned to the ALKYL-OH substructure. The final results of static/dynamic interpretation, in the form of a very concise, nonredundant list of substructures detected, are shown in Table V. We see that the algorithm features very high efficiency.

Clearly, both the network algorithms and the matrix algorithms (especially in the static/dynamic version) resemble the natural approach to interpretation of spectra, used by very experienced spectroscopists. We should add in this place some general comments: the network and matrix algorithms use only the knowledge on spectral-structure correlations, available throughout the literature. But, the comparison of various manuals and correlation tables shows how divergent information about spectral data (group frequencies, chemical shifts, etc.) is contained herein. Besides, both types of identification algorithms have three common features: (1) usually, only the substructures but not full structures may be detected;[58] (2) the substructures not active in a given spectral technique are wholly not recognized; (3) owing to the inherent ambiguity of the empirical interpretation of a spectrum (any type), as each band or peak may well represent various structural fragments, the algorithms discussed give distinct noise of information. This results in detection of substructures not present in the molecule analyzed (SRE, substructures recognized erroneously).

The general classification of identification algorithms presented here is one of convenience. Several known examples may belong simultaneously to more than one type. One can easily imagine, for instance, the structure determination by means of library search (as a preliminary step) and then the automatic application of network or matrix algorithms.

INTEGRATED SYSTEMS FOR STRUCTURE DETERMINATION: THEIR ADVANTAGES AND DISADVANTAGES

According to the definition given under Introduction, integrated systems for computer-aided structure determination use more than one spectral technique in the identification process and are able to supply the final results in the form of a structural formula[59] of the compound being tested. Their detailed structure and results of some example applications are described elsewhere.[5-18] Here, we are rather in a position to discuss the problem-solving strategies used throughout integrated systems. The overall strategy of performing tasks by those systems consists of five distinct steps:[48]

(1) Correlation. Inferring the possible structural fragments by empirical interpretation of some selected molecular spectra, from the series MS, ^{13}C NMR, ^{1}H NMR, IR, and UV (the identification algorithms applied here are of types discussed in the paragraph devoted to monosystems).

(2) Consistency Test. Selection from the set of structural fragments found in step 1 those substructures that are internally consistent. We should remember that the structural data gained by interpretation of several different spectra for the same molecule may confirm, complete, or mutually eliminate.

(3) Structural Assembly. Combination of substructures found in step 2 into a meaningful total structure (tentative or candidate structure).

(4) Spectrum Prediction. Prediction of selected spectral features (like, for instance, number of bands) for the candidate structure arrived in step 3.

(5) Spectra Comparison. Comparison of predicted and experimental spectra (or respective, selected features). If they agree, the candidate structure may be correct and is printed. If no, the candidate structure is erased from computer memory. For both cases, thereafter return to step 3 to generate another tentative structure; if no more candidate structures can be generated, the process ceases.

This general problem-solving method, executed by the computer, again reflects strongly the sequence of operations of the human analyst while determining an unknown organic structure. After interpretation of available spectra (step 1), two problems for the scientist still remain: (a) the determination of substructures not identified from spectral data and (b) the possible sequences in which the substructures may be attached. Both problems are solved by simultaneous inspection of results from different spectral techniques and/or some simple tips, like calculation of degree of unsaturation (from empirical formula), determination of residual fragments by substraction of the formulas (or formula weights) of all the unique known substructures from the formula (or molecular weight), etc. This is done by the computer in step 2. And also, when the computer generates structural formulas from substructures detected (step 3), the whole procedure models a chemist doing the same job: polyvalent substructures are combined first in all possible ways; then, monovalent groups are attached, also in all possible ways. Finally, the various possible structures are checked to see if they are consistent with all known information about the compound. In computer realization, this is done in step 4.

It may be assumed that further development of the systems discussed here depends strongly on improvement of the procedures for elimination of excessively recognized substructures (step 2) and for elimination of redundant candidate structures (step 5). However, the overall development of the systems and rise of their efficiency are also directly connected with the maturity of algorithms for correlation identification (step 1). Hence, the capability limit of the systems discussed may be clearly noticed, as the efficiency of the identification algorithms converge to the given level of saturation. It is hard to believe that someone may elaborate the correlation algorithm, giving in output exclusively the substructures recognized properly, without excessive, redundant information. Other drawbacks of the systems are connected with their internal design and with the organization of information flow. Namely, the systems allow one to apply some a priori fixed combination of spectral methods only (for instance, ^{1}H NMR, IR, and UV),[15] and they are not ready for application of all spectral methods used in routine analytical problems (MS, ^{13}C NMR, ^{1}H NMR, IR, Raman, UV). On the other hand, the idea of projecting automata has had mischievous influence on the internal structure of the systems: they perform their task automatically, almost without human intervention, up to giving the final printout with candidate structures. Although in some cases supplementary information could be entered by the analyst while the systems are running, it might be thought that here the human being assists the computer.

As a consequence of the described design of the systems, extremely large difficulties are discovered in the case of any improvement, say during the inclusion or elimination of given subroutines, change of the number substructures recognized, etc. This state bears a resemblance to properties of spaghetti: any trial to cut out even a very small piece entails the movement of the whole portion.

IIAI APPROACH TO PROBLEM SOLVING IN STRUCTURE DETERMINATION

Recently, the novel approach to computer-aided structure determination has been announced,[49] namely, the Independent Identification by Artificial-Intelligence Interpreters, IIAI, exemplified by the program environment called SCANSPEC.[60] The artificial-intelligence interpreters are defined as packages of utility programs for interpretation of MS, ^{13}C NMR, ^{1}H NMR, IR, RA, and UV spectra. These interpreters are fully independent of each other, but the results of their work are loaded onto common disk, for usage in further steps of the analytical process. The user has in this situation a very flexible choice of spectral method, according to any special requirements resulting from the experiment and available instrumentation. Then, structural assembly and elimination of improper candidate structures are executed by independent utility programs, stored on the same disk. Besides, the man–machine communication tools are so formulated that the computer truly assists the analyst, but not inversely.

CONCLUSIONS

The present status of the systems for computer-assisted structure determination may be briefly summarized as follows: the systems operate at the level of post doc analyst; their performance is good not because they know any more than an experienced spectroscopist but because (a) they use most of the rules applied by an analyst to solve structure determination problems, (b) they apply the same set of rules in every task, even in routine problems, and (c) they apply systematically the whole set of rules each time, without mistakes and loss of memory.

But in the near future, we may expect further improvement in automatic determination of structures, connected with development of theory (problem-solving methods, better and more reliable statistically spectrum–structure correlations) and of hardware (release from limitations in computer memory and machine time).

REFERENCES AND NOTES

(1) Ungan, S. J. *Pure Appl. Sci.* **1975**, *8*, 305.
(2) Dessy, R. E. *Anal. Chem.* **1984**, *56*, 1200A.
(3) Tormyshev, V. M.; Derendyaev, B. G.; Koptyug, V. A. *Anal. Chim. Acta, Comp. Techn. Opt.* **1981**, *133*, 517.
(4) Hippe, Z. *Anal. Chim. Acta* **1983**, *150*, 11.
(5) Delfino, A. B.; Buchs, A. *Fortschr. Chem. Forsch.* **1973**, *39*, 109.
(6) Michie, D.; Buchanan, G. B. In "Computer for Spectroscopists"; Carrington, R. A. G., Ed.; Hilger: London, 1974; p 114.
(7) Masinter, L. M.; Sridharan, N. S. In "Computer Representation and Manipulation of Chemical Information"; Wipke, W. T.; Heller, S. R.; Feldman, R. J., Eds.; Wiley: New York, 1974; p 287.
(8) Lindsay, R. K.; Buchanan, G. B.; Feigenbaum, E. A.; Lederberg, J. "Application of Artificial Intelligence for Organic Chemistry-The Dendral Project"; McGraw-Hill: New York, 1980.
(9) Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Crandell, C. W. *Anal. Chim. Acta* **1981**, *133*, 471.
(10) Shelley, C. A.; Woodruff, H. B.; Snelling, C. R.; Munk, M. E. In "Computer-Assisted Structure Elucidation"; Smith, D. H., Ed.; ACS: Washington, DC, 1977; p 92.
(11) Shelley, C. A.; Munk, M. E. *Anal. Chim. Acta* **1981**, *133*, 507.
(12) Yamasaki, T.; Abe, H.; Kudo, Y.; Sasaki, S. In "Computer-Assisted Structure Elucidation"; Smith, D. H., Ed.; ACS: Washington, DC, 1977; p 108.
(13) Sasaki, S.; Abe, H.; Fujiwara, I.; Yamasaki, T. In "Data Processing in Chemistry"; Hippe, Z., Ed.; PWN-Elsevier: Warsaw, 1981; p 186.
(14) Hippe, Z. Res. Memor. MR-I-32; Technical University Publications: Rzeszów, 1981.
(15) Debska, B.; Duliban, J.; Guzowska-Swider, B.; Hippe, Z. *Anal. Chim. Acta* **1981**, *133*, 303.
(16) Hippe, Z. In "Data for Science and Technology"; Glaeser, P. S., Ed.; North-Holland: Amsterdam, 1983; p 107.
(17) Gribov, L. A.; Elyashberg, M. E.; Serov, V. V. *Anal. Chim. Acta* **1977**, *95*, 75.
(18) Gribov, L. A. *Anal. Chim. Acta* **1980**, *122*, 249.
(19) Dubois, J. E.; Bonnet, J. C. *Anal. Chim. Acta* **1979**, *112*, 245.
(20) Heller, S. R. In "Data Processing in Chemistry"; Hippe, Z., Ed.; PWN-Elsevier: Warsaw, 1981.
(21) Koptyug, V. A. In "Computer Applications in Chemistry"; Heller, S. R.; Potenzone, R., Jr., Ed.; Elsevier: Amsterdam, 1983; p 207.
(22) Sasaki, S.; Abe, H.; Fujiwara, I.; Yamasaki, T.; Hippe, Z.; Debska, B.; Duliban, J.; Guzowska-Swider, B. *Chem. Anal. (Warsaw)* **1982**, *27*, 171.
(23) Bremser, W. *Chem.-Ztg.* **1982**, *104*, 53.
(24) Inose, H. "Research on Scientific Information Systems in Japan"; Computer Centre, University of Tokyo: Tokyo, 1980.
(25) Sebesta, R. W.; Johnson, G. G., Jr. *Anal. Chem.* **1972**, *44*, 260.
(26) Koptyug, V. A. *Z. Chem.* **1975**, *15*, 41.
(27) Wangen, L. E.; Woodward, W. S.; Isenhour, T. L. *Anal. Chem.* **1971**, *43*, 1605.
(28) Knock, K.; Venkataraghavan, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973**, *95*, 4185.
(29) Abramson, F. P. *Anal. Chem.* **1975**, *47*, 45.
(30) Kwiatkowski, J.; Riepe, W. *Anal. Chim. Acta, Comp. Techn. Opt.* **1979**, *112*, 219.
(31) Erley, D. S. *Appl. Spectrosc.* **1971**, *25*, 200.
(32) Razinger, M.; Penca, M.; Zupan, J.; Janezic, M. *Fresenius' Z. Anal. Chem.* **1982**, *313*, 496.
(33) Hippe, R.; Hippe, Z.; Fic, G. "Application of Computers in Processing of Physicochemical and Analytical Data"; Technical University Publications: Rzeszów, 1979; p 53.
(34) Munk, M. E.; Shelley, C. A.; Woodruff, H. B.; Trulson, M. O. *Fresenius' Z. Anal. Chem.* **1982**, *313*, 473.
(35) Hilsenrath, J. "Summary of On-Line or Interactive Physicochemical Numerical Data Systems"; National Bureau of Standards: Washington, DC, 1980.
(36) Creswell, C. J.; Runquist, O. A.; Campbell, M. M. "Spectral Analysis of Organic Compounds"; Longman: Edinburgh, 1972.
(37) Clerc, J. T.; Pretsch, E.; Seibl, J. "Structural Analysis of Organic Compounds by Combined Application of Spectroscopic Methods"; Elsevier: Amsterdam, 1981.
(38) Dabrowski, A. "On Theory of Information"; WSIP: Warsaw, 1974; in Polish.
(39) Mazurkiewicz, A. "Problems of Information Processing"; Wydawn. Nauk.-Tech.: Warsaw, 1974; in Polish.
(40) Hippe, Z.; Debska, B. *Bull. Acad. Pol. Sci., Ser. Sci. Chim.* **1974**, *22*, 551.
(41) Hippe, Z.; Kerste, A. *Bull. Acad. Pol. Sci., Ser. Sci. Chim.* **1974**, *22*, 541.
(42) Hippe, Z. "Application of Computers in Processing of Physicochemical and Analytical Data"; Technical University Publications: Rzeszów, 1979; p 43.
(43) Hippe, Z.; Hippe, R.; Duliban, J. *Fresenius' Z. Anal. Chem.* **1982**, *311*, 440.
(44) Hippe, Z. *TrAC, Trends Anal. Chem. (Pers. Ed.)* **1983**, *2*, 240.
(45) Hippe, Z.; Duliban, J.; Licbarska, R.; Kozioł, J.; Mazur, M. *Chem. Anal. (Warsaw)*, in press.
(46) Jamróz, M.; Latek, Z. *J. Mol. Struct.* **1984**, *115*, 277.
(47) Jamróz, M.; Latek, Z.; Hippe, Z. "Moderne Entwicklungsrichtungen der Molekülspektroskopie", Eisenach, Oct 1–5, 1984.
(48) Clerc, J. T.; Koenitzer, H. In "Data Processing in Chemistry"; Hippe, Z., Ed.; PWN-Elsevier: Warsaw, 1981; p 151.
(49) Hippe, Z. "EUROANALYSIS-V", Kraków, Aug 26–31, 1984.
(50) Entropy of information on the molecular structure, contained in a given spectrum, decreases distinctly in the sequence: MS, ^{13}C NMR, ^{1}H NMR, IR, RA, and UV.
(51) This general solution is applied, even if a given author(s) does (do) not mention it explicitly.
(52) Both types of searching (positive and negative) copy exactly the procedures used by a human being while he goes through the spectra manual.
(53) The algorithm discussed may be organized more effectively, using the method described by us.[33] Namely, the data base for search system is organized in special way; all absorption bands for each reference spectrum are first sorted according to intensities and just then stored onto disk. Therefore, the user may instruct the computer to compare the first two bands, the first three bands, the first four bands, etc. This procedure, very effective in practice, emphasizes the importance of most intense bands in structure determination.
(54) In practice, $\sim 10 < N \lesssim 400$, depending on the type of spectrum analyzed. The largest number of substructures detected is for ^{13}C NMR and ^{1}H NMR spectra.
(55) Full binary tree of order k has 2^k terminals.
(56) The matrix algorithms have been successfully applied also to other spectral techniques, for instance, ^{1}H NMR, ^{13}C NMR, UV, and RA.[45]

(57) Each substructure determined in the algorithm has its own decision function.
(58) The only exception known is the recognition of vinyl polymers and copolymers, described in reference 41.
(59) Provided the empirical formula is known explicitly or from MS.
(60) This research project is realized under the auspices of the Polish Academy of Sciences, as part of the problem MR-I-32, "New Analytical Methods".

Computer Systems for Laboratory Networks and High-Performance NMR

GEORGE C. LEVY* and JOHN H. BEGEMANN†

NIH Biotechnology Research Resource for Multi-Nuclei NMR and Data Processing, Syracuse University, Syracuse, New York 13210

Received January 1, 1985

Modern computer technology is significantly enhancing the associated tasks of spectroscopic data acquisition and data reduction and analysis. Distributed data processing techniques, particularly laboratory computer networking, are rapidly changing the scientist's ability to optimize results from complex experiments. Optimization of nuclear magnetic resonance spectroscopy (NMR) and magnetic resonance imaging (MRI) experimental results requires use of powerful, large-memory (virtual memory preferred) computers with integrated (and supported) high-speed links to magnetic resonance instrumentation. Laboratory architectures with larger computers, in order to extend data reduction capabilities, have facilitated the transition to NMR laboratory computer networking. Examples of a polymer microstructure analysis and in vivo ^{31}P metabolic analysis are given. This paper also discusses laboratory data processing trends anticipated over the next 5-10 years. Full networking of NMR laboratories is just now becoming a reality.

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy and (recently) magnetic resonance imaging (MRI) have shown remarkable development as powerful and increasingly influential analytical techniques.[1-3] In the early 1960s, NMR spectroscopy first received wide application to solution chemical structure elucidation; by 1985, NMR is ubiquitous in applications to solution and solid-state studies of small molecules, biological and synthetic polymers, molecular complexes, and even complex biosystems such as living cells.[1] Current NMR methods examine nuclei across the periodic table with increasingly complex multipulse experiments that can be designed to probe specific features of these structures. In magnetic resonance imaging, complex radio-frequency pulse sequences combine with pulsed magnetic field gradients to spatially encode resonance frequency information. Magnetic resonance images promise revolutionary change in diagnostic medicine.[3]

All of these modern NMR and MRI experiments are totally dependent on current computer techniques. In the 1960s, NMR spectrometers could add simple "computer" signal averagers; by the early 1970s, most research NMR instrumentation operated in pulse Fourier-transform mode, necessitating dedicated, if not fully integrated, minicomputer systems to control pulse generation and data acquisition, both at time resolutions of 1-100 μs. During the 1970s, commercial NMR spectrometers integrated their dedicated computers until by 1980 the computer was the central subsystem of most NMR instruments. A recent trend has been to utilize multiple computers in NMR instrumentation, distributing the tasks of pulse generation, data acquisition, monitoring instrument functions, and providing user interaction. In 1985, research NMR spectrometers utilize fast 20-32-bit word length mini-computers for central functions, with coordinated 8- or 16-bit microprocessors dedicated to spectrometer control and oversight tasks. Newly developing magnetic resonance imaging instrumentation, designed for clinical use, extends these trends with near state of the art computer and graphics technology.

Despite the advances in application of computer methods over the past 25 years, current NMR instrumentation has lagged in utilization of one of the most important developments in computer architecture: *networking*. Part of the reason for this is understandable. Until the mid 1980s, "universal" computer hardware interconnections were limited to slow protocols such as the RS-232 serial link. NMR data files, especially 2-dimensional data sets, can be extremely large—several megabytes or even much larger. NMR data transfers are thus very inefficient by these slow links.

Ethernet,[4] which is the current "standard" for local area networking, offers very high speed (10 million bits/s), but the Ethernet collision detection base-band design is not nearly ideal for large laboratory networks that can transfer very large files from several spectrometers. Unfortunately, implementation of useful NMR laboratory computer networks is currently restricted by the lack of widely implemented alternatives to Ethernet. Thus, heterogeneous configurations of spectrometers and other network nodes (fileservers for data storage, laboratory concentrators for off-line processing, mainframes if desired, etc.) mandate use of Ethernet, except when an alternative structure is provided by manufacturers—and currently this would limit networking to interconnecting instruments of individual manufacturers, without fully supported links to other computers.

There are several computer network activities especially relevant to NMR laboratories: (1) removal of primary data reduction from spectrometer computers to increase overall laboratory efficiency; (2) provision for archival data storage; (3) use of larger laboratory superminicomputers or mainframe computers to run sophisticated computational software for data reduction, simulations, etc.; (4) development of interinstrument and interlaboratory data sharing, necessitating a common implementation of data formats or "translators" for each format present on the network; (5) eventual integration of NMR data processing into *comprehensive* Laboratory Information Management Systems (LIMS).

† Present address: New Methods Research, Inc., Jamesville, NY 13078.

0095-2338/85/1625-0350$01.50/0 © 1985 American Chemical Society