Pure & Appl. Chem., Vol. 50, pp. 103-106. Pergamon Press, 1978. Printed in Great Britain.

COMPUTER METHODS FOR THE SPECTROSCOPIC IDENTIFICATION OF ORGANIC COMPOUNDS

J.T. Clerc

Laboratorium für Organische Chemie, Eidgenössische Technische Hochschule
Zürich, CH-8092 Zürich, Switzerland

Abstract - Various spectroscopic methods are used for the identification of organic compounds. The most commonly used methods include infrared spectroscopy, ultraviolet/visible spectroscopy, nuclear magnetic resonance, and mass spectrometry. For the interpretation of the spectroscopic results semi-empirical methods are generally used. These methods of interpretation are based on large numbers of previously recorded reference spectra. The reference data compilations are used either to find semi-empirical correlation rules (correlation charts) or for direct comparison with the spectral data of an unknown sample. In either case, large amounts of data have to be processed. Human beings are generally not very good at shuffling large amounts of data, but are capable of detecting very complex and sophisticated relationships. On the other hand, computers are extremely good at doing simple routine jobs very quickly and virtually free from error. Thus for the interpretation of spectroscopic data the best performance can be expected from a "man-machine combo", where both partners do their special parts of the job.

During the last half-century, tremendous progress has been made in the elucidation of the structure of organic compounds. To illustrate this let us arbitrarily select as an example the first organic compound listed in the catalogue of a supplier of fine chemicals (1), namely abietic acid. A Ph.D. thesis (2) dated 1935 and prepared at the ETH Zürich had as one of its main topics the determination of the number and position of the carbon-carbon double bonds in abietic acid. In this excellent work the author was able to reduce the considerable number of possibilities for the isomers to 18. They are depicted in Fig.1.

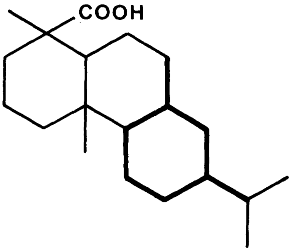

Fig. 1. Possible isomers for abietic acid [from a Ph.D. thesis (2)]. Two of the heavy-lined bonds are double bonds. There are 18 possibilities if cumulated double bonds are excluded.

However, the methods available at that time virtually precluded further progress. If we were confronted with the same problem today, various spectroscopic methods would be used to eliminate most of the remaining tentative structures.

To begin with, we record the ultraviolet spectrum of abietic acid dissolved in ethanol. From the curve we can read the wavelength for the absorption maximum corresponding to the $\pi \rightarrow \pi^*$ transition, namely 240 nm. The extinction coefficient is calculated to be 1.55×10^4 $l.mole^{-1}$ cm^{-1}. Comparing these values with reference data we can conclude that the two double bonds must be conjugated. Given this information we can now estimate the expected values for the $\pi \rightarrow \pi^*$ transition for all possible isomers, with good accuracy. Application of the relevant rules [cf. (3)] indicates that those structures having both double bonds in the same ring would exhibit an absorption maximum at a wavelength considerably longer than 255 nm and are therefore at variance with the experimental values. Thus a simple ultraviolet spectrum rules out at once 14 of the 18 isomers we started with. Of the remaining 4 possibilities, 3 may be eliminated with the help of other spectroscopic methods, leaving us with the correct solution for the unknown structure. The measurements as well as the interpretation of the spectra can easily be done in one working day; thus a problem which 40 years ago formed the better part of a top quality Ph.D. thesis can now be rapidly solved.

Let us now analyse what we did in this very simple, but rather typical, example. First, we did some preprocessing on the recorded ultraviolet spectrum by reading off the wavelength and calculating the extinction coefficient for the $\pi \rightarrow \pi^*$transition. By correlating these values with reference data we concluded that the two double bonds must be conjugated. For all tentative structures we then made a prediction of the wavelength of the absorption maximum. Comparing these values with one actually measured permitted exclusion of most of the structures considered, leaving us with a comparatively small number of possible structures. We can hence easily identify the five basic steps performed in the elucidation of the structure of organic compounds by spectroscopic methods. They are summarized in Table 1.

TABLE 1. Processing steps in structure elucidation by spectroscopic methods

Step 1 Preprocessing

Eliminate from raw data all features that are irrelevant and/or confusing. Transform data into a form easy to process.

Step 2 Correlation

Identify structural units which must be (a) present or (b) absent in the unknown structure, using known correlations between spectral and structural features as well as any non-spectroscopic information available.

Step 3 Tentative structures

Assemble tentative (partial) structures which are in agreement with the structural elements found in step 2.

Step 4 Spectrum prediction

Predict spectral data for the proposed (partial) structures, using correlation tables, additivity rules, reference spectra, theoretical calculations, etc.

Step 5 Comparison of spectra

Compare the predicted spectrum with the experimental one. If they agree within reasonable error limits, the proposed structure may be correct. If they do not agree, modify the tentative structure so as to get a better fit, and repeat steps 4 and 5.

For a fully automated system of organic structure elucidation all five steps would have to be performed automatically. To get some idea of the type and order of magnitude of the problems involved in implementing such a system, let us take a closer look at these five basic operations.

Preprocessing involves data reduction as well as various data transformations. The main motive for the latter is to obtain data representations which are more easily processed and interpreted. A good example is given by the Fourier transformation of ^{13}C-NMR interferograms from the time domain into the frequency domain. The transform no longer contains the irrelevant (and often confusing) phase information and is well suited for correlation with reference data. The methods applied in the preprocessing step depend very much on what type of further processing of the data is intended. Fast, efficient and sophisticated preprocessing implies that the complete set of raw data is available in form readable by computer. In many cases this is a prerequisite not easy to implement in a routine analytical laboratory. However, computerized data acquisition is not a topic to be discussed in this lecture.

Correlation of the spectral data with structural features involves pattern recognition. One of the basic problems encountered is that human beings are such extremely efficient pattern recognizers. For us to recognize the face of a friend in a crowd poses no problem. The design of a computer program that accomplishes the same task is virtually impossible, even with today's most advanced and sophisticated hard- and software. Similarly, an experienced analyst will easily identify various functional groups and substructural units from spectroscopic data by making use of his pattern-recognition abilities. He will use extremely sophisticated decision criteria, probably too complex to be concisely expressed even in his own words. Thus, all attempts to use mathematical methods of pattern recognition published to date have either been only moderately successful at best, or are restricted to those classes of very simple compounds which exhibit an outstandingly simple and well understood spectroscopic behaviour. In summary, we can state that mathematical methods of pattern recognition work well on those compounds which do not present problems to the analyst anyway. i.e. where there is no urgent need for computer-aided interpretation of spectra.

Assembling all possible tentative structures from a given set of substructural elements is a job which the computer can definitely do faster and better than a human. The computer, if properly programmed, will generate all possible structures in an unbiased way. However, the necessary programs are involved and the run-time tends to grow almost exponentially with the number of substructural elements under consideration. There exists also a quantitative problem, as the number of different permutations of n distinct elements is $n!$ and thus the number of admissible building blocks has to be drastically curbed (the age of the universe is about 19! seconds) and/or conditions restricting the set of allowable combinations have to be built into the program.

Spectrum prediction is relatively easy for very simple model compounds. For those compounds encountered in the analyst's real world, however, the necessary calculations are prohibitively expensive and complicated, if possible at all. Furthermore, the accuracy attained is often rather modest. This has to be taken into account when comparing spectra. The comparison of two spectra has to be performed with a fair amount of tolerance to allow for deviations arising from poor estimates as well as from various instrumental and technical artifacts.

Several approaches to the automation of the complete system have been published [cf. e.g. (4)-(9)]. They give surprisingly good results when applied to simple model compounds belonging to the compound classes the systems are designed for. For the solution of real-world problems, however, their performance is feeble. This limits the successful application of fully automated systems today to problems where the interpretation of spectra is straightforward or even trivial. Putting off the development of such systems as being useless intellectual play is certainly not appropriate, though. Research in this field has significantly increased our understanding of the methods used in semi-empirical spectroscopy and has led to valuable contributions to many other branches of science and technology.

Another approach to the computer-aided identification of organic compounds is the library-search method. Here, the spectrum of the unknown is compared with all entries in a library of reference spectra. If a perfect match is found between the spectral data for the unknown and a reference compound, it is assumed that the structures will also be identical. This seemingly simple approach fits well into the general pattern of structure elucidation by spectroscopic methods as described in the foregoing section. We may identify the same five basic steps, as depicted in Table 2. The problems associated with steps 2-4 are now replaced by the problem of providing a suitable library of reference spectra.

TABLE 2. Basic processing steps (cf. Table 1) applied to library search

Step 1	Preprocessing: for fast and easy comparison
Step 2	Correlation: not done
Step 3	Tentative structures: use all structures in the library
Step 4	Spectrum prediction: use reference spectra from the library
Step 5	Comparison of spectra: to retrieve compounds with similar (as opposed to identical) structures

Even with very large libraries the probability of finding an exact match is extremely low, and for new compounds it is always zero. Thus, the comparison step has to be implemented in such a way that retrieval of similar (as opposed to identical) structures becomes possible. We no longer search for an exact match, but rather for the entry most similar to our model. The crucial point is to avoid comparing exact values for single spectral parameters but rather to rely on measures describing the general pattern of a given spectrum. In mass spectrometry, for example, very precise intensity data for peaks at selected m/z values should be used with caution. More complex measures, such as the general distribution of the ion abundances over the mass range or the relative intensities of ion series, are to be preferred (10). In ^{13}C-NMR spectroscopy we rely more on the relative number of peaks in relatively wide shift ranges than on the exact shift values for single peaks. The proton/carbon ratio, which may easily be evaluated from the number of peaks in the off-resonance decoupled spectrum has also proved to be useful (11). A library-search system using features of this type for the comparison step is well suited for retrieving from the library compounds with structures similar to the unknown, e.g. homologous compounds, compounds with the same skeleton, or derivatives. It will therefore supply useful answers even if the library does not contain a reference compound identical to the unknown.

As the spectrum of the unknown has to be compared with all spectra in the library, the run-time and operating costs of a library-search system are dominated by the time needed for the comparison step. Therefore, for an economically feasible system the comparison step should be as fast as possible. This calls for encoding the spectral data in binary variables of the yes/no type. It is the task of the preprocessing step to provide for this encoding. Preprocessing of the reference library is performed only once, when the system is initially set up. The comparison step, however, is performed many times, once for each library compound. It is therefore advantageous to optimize the system for highest comparison speed, even at the expense of a complex and relatively time-consuming encoding algorithm. The computing time will still be dominated by the comparison step.

Library-search systems have one inherent weak point. If the library does not contain a suitable reference compound for a given unknown, the result will be useless, no matter how complex and sophisticated the program is. Thus, the quality of the reference library is one of the key factors governing the performance. We believe that today's library-search systems are much more limited by the libraries they use than by the encoding and comparison algorithms. Therefore, we feel that for the time being, highest priority should be given to the compilation of suitable libraries of spectra (12). In order to provide the system with suitable reference data for a broad range of problems with a library of limited size, we should primarily select for inclusion simple model compounds rather than highly complex molecules. Furthermore, large series of homologous compounds may be reduced to a few representative examples without significantly affecting the utility of the data collection. The size of the compilation is of only secondary importance. Heavy emphasis should rather be placed on an unbiased representation of as many compound classes as possible, with complete sets of carefully verified data.

REFERENCES

1. Fluka Catalogue (1975).
2. H. Waldmann, Dissertation, ETH, Zürich (1935).
3. E. Pretsch, J.T. Clerc, J. Seibl and W. Simon, Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden, p. U30, Springer, Berlin (1976).
4. H. Abe and S. Sasaki, Sci. Rept. Tokoku Univ., Ser. 1 55, 63 (1972).
5. N.A. Gray, Anal. Chem. 47, 2426 (1975).
6. R.E. Carhart, D.H. Smith, H. Brown and C. Djerassi, J. Am. Chem. Soc. 97, 5755 (1975).
7. H. Rotter and K. Varmuza, Anal. Chim. Acta, Comp. Techn. Opt. 95, 25 (1977).
8. H.B. Woodruff and M.E. Munk, Anal. Chim. Acta, Comp. Techn. Opt. 95, 13 (1977).
9. L.A. Gribov, M.E.E. Elyashberg and V.V. Serov, Anal. Chim. Acta, Comp. Techn. Opt. in the press.
10. P.R. Nägeli and J.T. Clerc, Anal. Chem. 45, 739A (1974).
11. R. Schwarzenbach, J. Meili, H. Könitzer and J.T. Clerc, Org. Mag. Res. 8, 11 (1976).
12. R. Büchi, J.T. Clerc, Ch. Jost, H. Könitzer and D. Wegmann, Anal. Chim. Acta, Comp. Techn. Opt. in the press.