Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with Their Mass Spectra

Bernhard Seebass and Ernö Pretsch*

Department of Organic Chemistry, Swiss Federal Institute of Technology (ETH), Universitätstrasse 16,
CH-8092 Zürich, Switzerland

Received December 30, 1998

A new algorithm has been designed to check the compatibility of a mass spectrum with a proposed molecular formula or structure. Peaks are regarded as valid if the corresponding mass agrees with an allowed elemental composition and the required isotope signals are present. In addition, connectivity information is taken into account to some extent. Although reasonable intensity tolerances are applied, numerous errors are normally found in databases, in most cases due to missing or inadequate isotope signals. Besides database checks, the program can also be used to automatically detect spectra that contradict the proposed molecular formula, e.g., in combinatorial analysis.

1. INTRODUCTION

One of today's challenges in spectroscopic analysis is the need to increase the efficiency of solving spectra−structure issues. There is, therefore, an urgent demand for automation in this field. In many cases, a proposed structure is available in electronic form along with the measured spectrum so that the question to answer is whether the two are in agreement. For compatibility tests in general, spectral features are predicted from the molecular structure and compared with the corresponding ones gained from the measured spectra. In ^{1}H and ^{13}C NMR spectroscopy, chemical shifts can be predicted with good reliability from a given structure[1−4] and then compared with the measured ones obtained automatically from the spectra. Various methods have been proposed to predict IR spectra by using a reference database and uniform-length structure descriptors.[5−8] The so-obtained spectra can then be directly compared with the measured ones. However, no such approach is applicable to mass spectra because, despite various efforts,[9] it has not been possible to reliably predict them. Therefore, computer programs developed for MS database checks[10,11] focus mainly on molecular ions and formal aspects. Although even the simple analysis of the number of signals seems to be informative,[12,13] procedures making more efficient use of the entire spectra are desirable.

This paper presents a novel procedure based on the simple idea of computing a constrained list of all possible elemental compositions from a given molecular formula. Peaks at masses that do not occur in this list or are not accompanied by the required isotope signals contradict the molecular formula proposed. Additionally, the list of elemental compositions can be further restricted by considering certain structural information. This not only enhances the power of the method but also allows the use of mass spectral data to constrain, to some extent, the list of possible isomers obtained from a structure generator.[14]

2. EXPERIMENTAL SECTION

The programs are written in ANSI C and have been developed on a Power Macintosh 8500 using the development environment CodeWarrior (Metrowerks Inc., 9801 Metric Blvd., Austin, TX 78758). Two EI mass spectra collections have been used. The first one, DB1, is the SpecInfo MS database[15] consisting of the carefully checked collections of Dr. D. Henneberg (Max Planck Institute, D-4330 Mühlheim) and Prof. J. Seibl (ETH, CH-8092 Zürich). The spectra of the second database, DB2, are unchecked. Entries containing elements that are not considered (transition metals, lanthanides, actinides, and noble gases) as well as those having three-center bonds were removed from the data sets. There remained 24 732 and 2 661 entries in DB1 and DB2, respectively. The data shown in this paper are derived from DB1 unless stated otherwise.

3. PROCEDURES

The compatibility test procedure is outlined in the flowchart of Figure 1. To begin with, the structure information is read from a standard format file and is checked as follows. Molecules outside the scope of the program (see Experimental Section) are rejected. Hydrogen atoms, usually omitted in structure inputs, are added assuming that every atom occurs with its minimal valence. Otherwise, they must be specified in the input. Subsequently, several consistency checks are performed making sure that the molecular formula corresponds to an uncharged, even-electron molecule with appropriate valences for all atoms.

For the validated molecule, the elemental compositions of possible fragments are generated and stored in a fragment list. On the basis of the molecular formula alone, each combination of available elements could, in principle, represent such a fragment. However, since a full list of all combinations would include unreasonable fragments and thus account for signals not originating from the molecule, it is constrained according to various criteria as explained in the following. First, the number of double bond equivalents

10.1021/ci980171b CCC: $18.00
Published on Web 06/05/1999

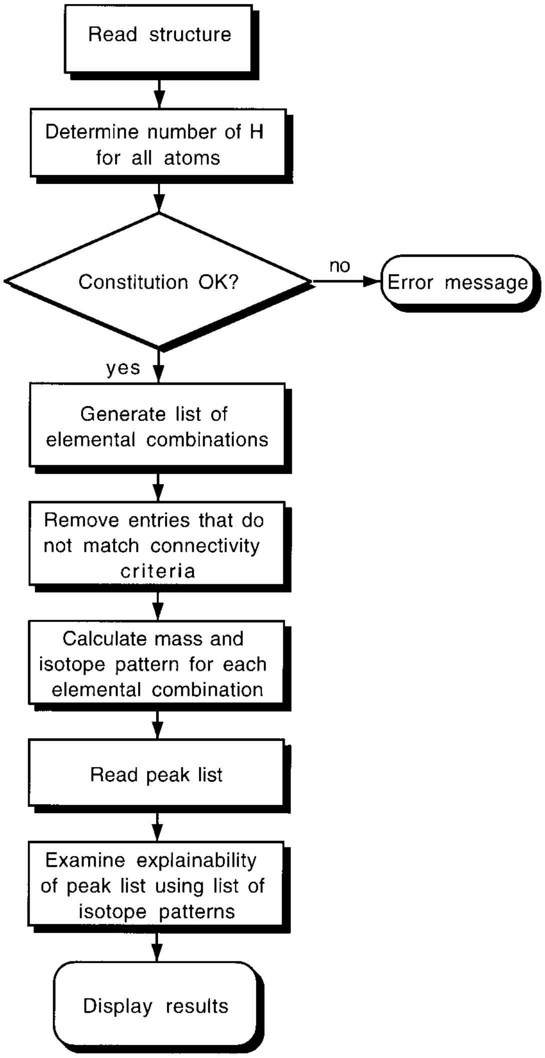
Read structure
Determine number of H for all atoms
Constitution OK?
no
Error message
yes
Generate list of elemental combinations
Remove entries that do not match connectivity criteria
Calculate mass and isotope pattern for each elemental combination
Read peak list
Examine explainability of peak list using list of isotope patterns
Display results

Figure 1. Flowchart of the basic procedure for the compatibility test of one structure and one mass spectrum.

(DBE) is limited. The lowest possible value of −0.5 admitted corresponds to a protonated saturated fragment or protonated molecular ion of a saturated compound. Except for the molecular ion, which may be protonated, the fragments must not have more H atoms than the molecule. Of course, the more DBE are allowed, the more signals will be explained, although increasingly by assuming unreasonable elemental combinations (e.g., C_3 for m/z 36). Therefore, an upper limit is set for the number of DBE relative to that defined by the molecular formula. Its value of +3.5 has been chosen by studying the influence of the number of DBE on the explained part of those mass spectra in DB1 that are fully accounted for when no upper limit of DBE is set. In Figure 2 (top), the mean relative total ion current (TIC) of the unexplained part of the spectra is plotted as a function of ΔDBE, i.e., of the maximal allowed increase in the number of DBE of a fragment relative to that of the molecule (cf. detailed procedure below). As expected, by augmenting ΔDBE, those parts of the spectra that cannot be explained from the elemental compositions steadily decrease. For a more detailed view, Figure 2 (bottom) shows the improvement achieved upon increasing ΔDBE in steps of 0.5. Apparently, the improvement is more pronounced when ΔDBE is increased to a half number than to a whole one.

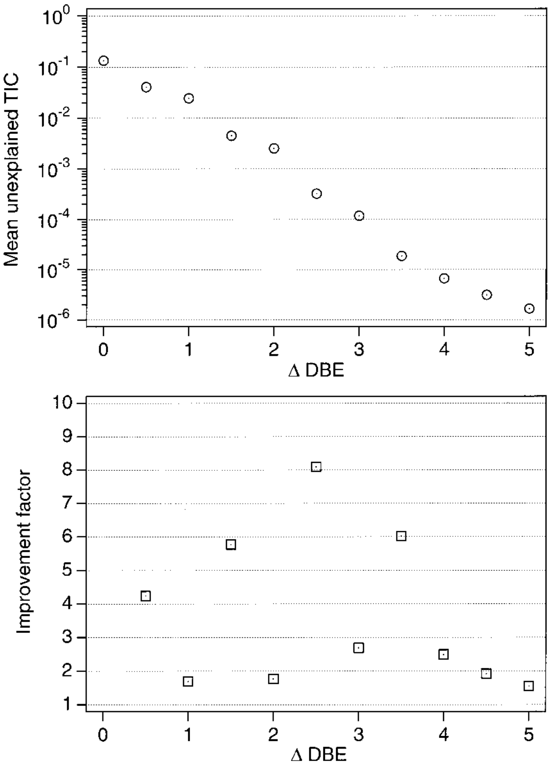

Figure 2. Top: mean unexplained total ion current (TIC) as a function of ΔDBE, i.e., of the maximal accepted difference in double bond equivalents (DBE) for a molecule and one of its fragments. Bottom: relative reduction in mean unexplained TIC when increasing ΔDBE in steps of 0.5.

This behavior can be explained by the fact that fragments with whole numbers of DBE, which correspond to radical cations, are less stable than those with half numbers, which correspond to the more stable even-electron species. For ΔDBE > 3.5, the improvement is small and for ΔDBE = 4.5, it is less than for ΔDBE = 4. This suggests that these improvements must be looked upon as artifacts in that they do not reflect real fragmentation processes but are due to unreasonable explanations of erroneous signals in the mass spectra.

In the next step, for each elemental combination of the thus restricted fragment list, the isotope signal intensities are calculated[16] and stored in a so-called pattern list sorted according to the mass of the most intensive isotope peak. In case this is the same for several of the patterns obtained, they are combined into one by calculating a minimal and a maximal intensity of each isotope peak, as shown in Figure 3, to later allow an efficient comparison with the measured spectrum.

Now, the experimental mass spectrum (for an artificial example, cf. Figure 4, top) is read in as a peak list. To account for tolerable inaccuracies, corresponding fuzzy intensities are calculated for each peak (Figure 4, center) by setting the minimal values to the measured ones, whereas maximal values are obtained by adding to the measured ones either 1% absolute or 10% relative, whichever is the greater.[17] Furthermore, minimal and maximal values of 0 and 1% are set if a specific mass has no entry. The two lists of fuzzy values, i.e., the combined pattern list from Figure 3 (bottom) and the peak list from Figure 4 (center), are then compared by the following heuristic procedure, which fully explains

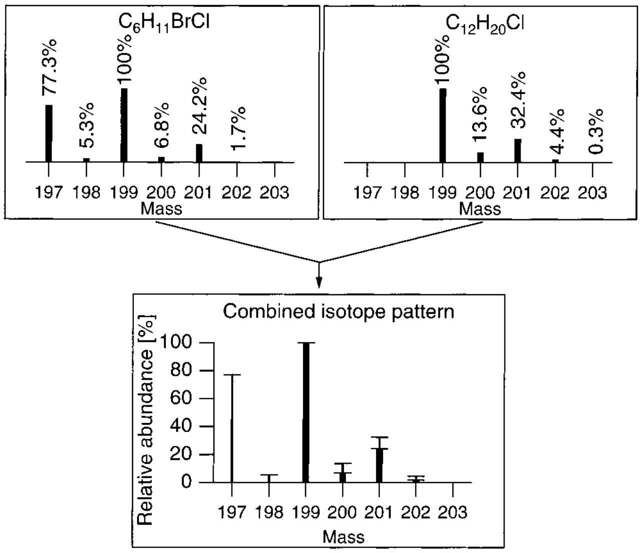

Figure 3. Top: two calculated isotope patterns. Bottom: representation of the corresponding fuzzy pattern combining the two isotope patterns.

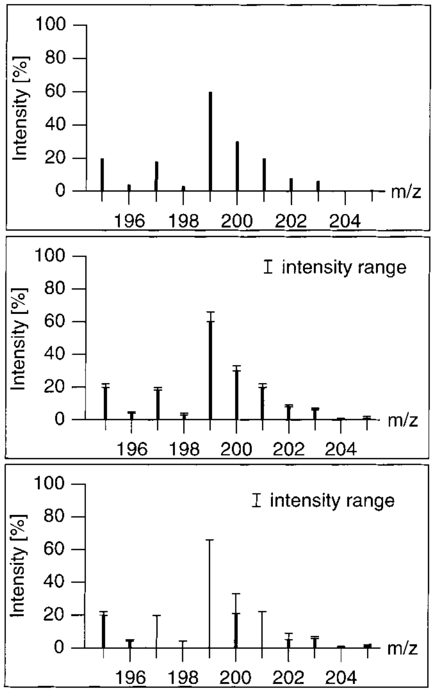

Figure 4. Part of an artificial mass spectrum. Top: raw data. Center: representation of the corresponding calculated fuzzy peak list. Bottom: minimal and maximal intensities obtained after comparing the fuzzy peak list (cf. center) with the calculated fuzzy isotope pattern from Figure 3 (bottom).

all valid peaks but may tolerate some invalid ones. Using minimal intensities from the calculated isotope pattern and maximal ones from the fuzzified measured spectrum, a multiplication factor (between 0 and 1.1) is calculated for each pattern of the pattern list that shows how often the isotope pattern fits in the spectrum. Then, for every mass, the minimal signal intensity is reduced by subtracting the product of this factor with the maximal intensity from the isotope pattern. Maximal intensities remain unaltered in order to accommodate the signals of neighboring fragments with their isotopes, if present (Figure 4, bottom). Ideally, at the end of this procedure, all minimal intensities of the peak list should equal zero. To admit the possibility of a weak background, only signals of intensity >1% are summed up to give the unexplained part of the mass spectrum. The procedure is illustrated by the example in Figure 4: The most intensive peak (*m*/*z* 199) is completely accounted for by the combined isotope distributions (Figure 3). To achieve this, the presence of a sufficiently intensive peak at *m*/*z* 201 is a prerequisite, hence, missing isotope signals can thus be detected. On the other hand, although the signal at *m*/*z* 197 is fully explained too, its unaltered maximal intensity would allow accommodation of the isotope signal of, e.g., a fragment containing Br at *m*/*z* 195.

So far, the procedure admits any elemental combinations compatible with the DBE constraint irrespective of connectivities. One could assume that in the absence of rearrangements (see below), only such fragments would occur that already exist as parts of the molecule. However, an exhaustive test would be prohibitively time-consuming. Therefore, as a computationally feasible compromise, the following connectivity-based rules have been implemented. The molecular structure is considered as composed of segments containing C atoms (C segments) on the one side and heteroatoms on the other, the individual C segments being separated by heteroatoms. Elemental compositions of the generated fragments are accepted only if the following conditions are fulfilled:

1. A fragment without heteroatoms must not contain more C atoms than the largest C segment.

2. A fragment with one heteroatom must not contain more C atoms than the C segments directly attached to that heteroatom.

3. For fragments with two or more heteroatoms, all atoms present in at least one path connecting any combination of two heteroatoms must also be present in the fragment.

For the first criterion to be applicable, the molecule must contain at least one heteroatom and at least two heteroatoms for the remaining criteria. All three restrictions ignore the possibility of rearrangements so that they may refuse valid elemental combinations. As an example, it is known that nitro compounds through rearrangement may eliminate NO. The corresponding product would not be valid according to the procedure outlined above because for nitro compounds criterion 2 forbids all fragments having C and O and no N. As a solution to this problem, common rearrangement reactions defined in an input file are considered in addition.

4. RESULTS AND DISCUSSION

The procedures developed in this work generate a list of allowed *m*/*z* values based on the possible elemental compositions which is constrained by the allowed number of DBE and, in case of known structures, by connectivity information. Using the conservative approach as outlined above, a spectrum should be fully explained by the correct constitution. Unfortunately, in most available databases, doubly charged ions are not correctly represented because only integer numbers of masses are stored. If such ions are present, their masses or the masses of some corresponding isotope peaks are odd-numbered, leading to half-numbered *m*/*z* values that are either ignored or stored under the next integer number. The algorithm described here correctly detects such wrong signals as errors. To circumvent this, the test modules can be constrained to the upper half of the mass spectra.

Table 1. Number of Database Entries and Some Characteristics for Four Selected Relative Molecular Masses[a]

	no. of entries					
M_r	total	valid	valid with heteroatom(s)	range of elements	range of DBE	no. of MF
100	96	84	33	$C_{2-7}H_{0-16}Cl_{0-1}F_{0-4}N_{0-2}O_{0-3}P_{0-1}S_{0-1}$	0−4	33
204	215	187	181	$C_{4-16}H_{0-25}F_{0-2}N_{0-6}O_{0-6}P_{0-1}S_{0-2}$	0−13	636
304	69	61	59	$C_{10-24}H_{8-40}Br_{0-1}Cl_{0-4}I_{0-1}N_{0-4}O_{0-8}P_{0-2}S_{0-4}$	0−18	701
392	39	33	33	$C_{12-31}H_{10-49}Br_{0-1}Cl_{0-2}N_{0-6}O_{0-10}P_{0-1}S_{0-4}$	0−22	1809

[a] DBE: double bond equivalents. MF: molecular formula.

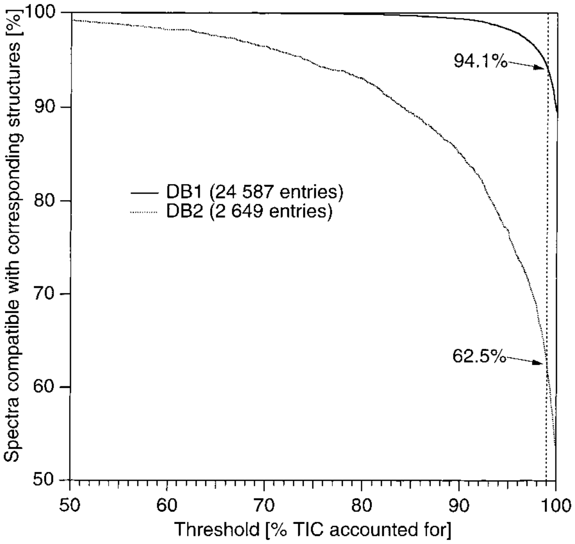

Figure 5. Compatibility test on two databases (DB1 and DB2) using molecular formula information only. Tolerances of intensity: 1% absolute or 10% relative (see text).

Tests on these lines show that similar results are obtained as when using the full spectra. To account for such effects, in this work 1% of the total sum of intensities of signals that cannot be explained is tolerated.

The performance of the procedures is first tested with both databases, DB1 and DB2 (see Experimental Section), without using the structural information. In Figure 5, the relative number of spectra is shown as a function of the sum of intensities (% TIC) of those peaks that are not explainable by this method. It shows that 5.9% of the mass spectra from the high-quality DB1 and 37.5% from DB2 have a summed intensity of >1% of unexplained signals although background signals of up to 1% are tolerated for each *m*/*z* value. By also considering the connectivity information, the number of refused spectra from DB1 and DB2 is increased to 6.3 and 41.8%, respectively (not shown). Most of the errors originate from incorrect isotope signal intensities as shown by the following experiment: If the intensity tolerance is increased from 1 to 10% absolute, the relative number of spectra found to be faulty decreases to 1.4 and 4.0% for DB1 and DB2, respectively. The results show that the procedure is suitable for the quality control of MS databases.

The additional applicability of the method for detecting invalid molecular formulas is tested as follows. All molecules containing no other elements than C, H, Br, Cl, F, I, N, O, P, and/or S were selected from DB1 for four chosen molecular masses (M_r = 100, 204, 304, and 392; cf. Table 1, columns 1, 2). Those spectra that violated the above tests were examined, and clearly erroneous ones were discarded. For the remaining ones (column 3), the range of elements (column 5) and the number of DBE (column 6) were determined. Using these restrictions, all possible molecular formulas were generated[18] (column 7), and each spectrum was tested against all molecular formulas of the corresponding mass. The mean percentages of incompatible molecular formulas are shown in Figure 6 for three different tolerances (5, 1, and 0%) in TIC. It can be seen that for a tolerance of 1%, the relative number of incompatible molecular formulas lies between 31 and 57% and decreases with increasing molecular mass. Although the method alone is not capable of determining the molecular formula of a compound from its mass spectrum, it seems to be of real use as a future component of automatic molecular formula generators using various spectra.[19−21]

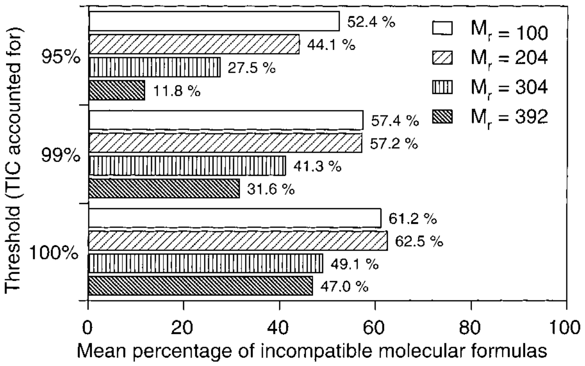

Figure 6. Compatibility test of generated molecular formulas and experimental spectra for molecules having the same relative mass.

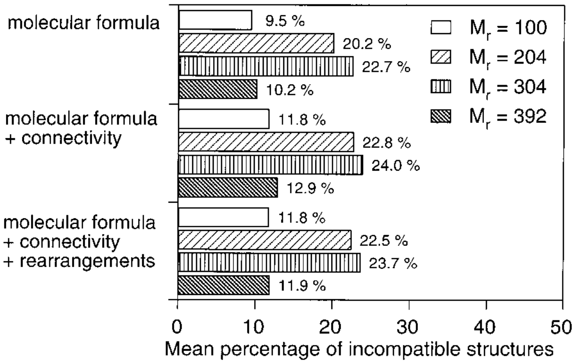

Figure 7. Effect of using structural information on the explained part of the spectrum when database entries of the same relative molecular mass are cross-checked.

Those compounds in Table 1 with at least one heteroatom (see column 4) were selected to test the usefulness of the connectivity information. Each spectrum was checked against all structures having the same molecular mass. For the threshold of 1% tolerance in TIC, the number of invalid entries increased by 2−3% if the connectivity information was considered in addition to the molecular formula (Figure 7). This number is only slightly smaller when rearrangements are also taken into account. The results indicate that mass spectrometry can be of some use for ranking and constraining the output of structure generators. Preliminary tests with 1 000–30 000 structures generated for four different molecular formulas show, however, that in most cases not more than a few percent of isomers can be excluded on the basis of their mass spectra. Actual work aims at further improving this figure.

In sum, our results show that useful tools are obtained by implementing highly conservative procedures that do not rely on any assumptions of fragmentation processes. They can be used (1) to check an entire database or new database entries, (2) to check a molecular formula against the corresponding mass spectrum, and (3) to constrain, to some extent, the output of structure generators. This last application is currently being improved. Although only EI mass spectra have been used here, the method does not imply any assumptions in this respect. It is, therefore, expected that it can also be used for compatibility checks in MS/MS applications.

5. CONCLUSIONS

The conservative procedure described for checking the possible occurrence of fragments on the basis of their elemental composition has proved to be efficient at detecting errors and eliminating invalid molecular formulas. The additional use of connectivity information leads to slight improvements. The program can be operated automatically using as input (1) a single structure and one mass spectrum, (2) a list of any number of structure–spectrum pairs, and (3) a series of structures and a single mass spectrum. The corresponding operation modes are designed (1) for automated analysis or database entry checks, (2) for database quality control, and (3) for constraining the output of a structure generator.

ACKNOWLEDGMENT

We thank Dr. R. Neudert (Chemical Concepts, D-69442 Weinheim) for providing the mass spectra, Dr. S. Stein (NIST, Gaithersburg, MD) and Dr. G. Bovermann (Novartis, Basel) for helpful discussions, and Dr. D. Wegmann for carefully reading the manuscript. Financial support by the Swiss National Science Foundation, Novartis Pharma AG (Basel), Hitachi Ltd. (Tokyo), and Orion Research Inc. (Beverly, MA) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Bremser, W. HOSE – a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355–365.

(2) Fürst, A.; Pretsch, E. A computer program for the prediction of ^{13}C NMR chemical shifts of organic compounds. *Anal. Chim. Acta* **1990**, *229*, 17–25.

(3) Chen, L.; Robien, W. OPSI: A universal method for prediction of ^{13}C NMR spectra based on optimized additivity models. *Anal. Chem.* **1993**, *65*, 2282–7.

(4) Bürgin Schaller, R.; Munk, M. E.; Pretsch, E. Spectra estimation for computer-aided structure determination. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 239–243.

(5) Affolter, C.; Clerc, J. T. Prediction of infrared spectra from chemical structures of organic compounds using neural networks. *Chemom. Intell. Lab. Syst.* **1993**, *21*, 151–157.

(6) Herges, R.; Weigel, U. M. Simulation of infrared spectra using artificial neural networks based on semiempirical and empirical data. *Anal. Chim. Acta* **1996**, *331*, 63–74.

(7) Baumann, K.; Clerc, J. T. Computer-assisted IR spectra prediction – linked similarity searches for structures and spectra. *Anal. Chim. Acta* **1997**, *348*, 327–343.

(8) Schuur, J.; Gasteiger, J. Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation. *Anal. Chem.* **1997**, *69*, 2398–2405.

(9) Gasteiger, J.; Hanebeck, W.; Schulz, K.-P. Prediction of mass spectra from Structural Information. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 264–271.

(10) Milne, G. W. A.; Budde, W. L.; Heller, S. R.; Martinsen, D. P.; Oldham, R. G. Quality control and evaluation of mass spectra. *Org. Mass Spectrom.* **1982**, *17*, 547–552.

(11) Terwilliger, D. T.; Behbehani, A. L.; Ireland, J. C.; Budde, W. L. The status and evaluation of a mass spectral database. *Biomed. Environ. Mass Spectrom.* **1987**, *14*, 263–270.

(12) McLafferty, F. W.; Stauffer, D. B.; Loh, S. Y. Comparative evaluation of mass spectral databases. *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 438–440.

(13) Stein, S. E.; Ausloos, P.; Lias, S. G. Comparative evaluation of mass spectral databases. *J. Am. Soc. Mass Spectrom.* **1991**, *2*, 441–443.

(14) Munk, M. E. Computer-based structure determination: Then and now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997–1009.

(15) SpecInfo, Chemical Concepts GmbH, P.O. Box 100202, D-69442 Weinheim. Germany.

(16) Kubinyi, H. Calculation of isotope distributions in mass spectrometry. A trivial solution for a nontrivial problem. *Anal. Chim. Acta* **1991**, *247*, 107–119.

(17) McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra*, 4th ed.; University Science Books: Mill Valley, 1993.

(18) Gloor, A.; Cadisch, M.; Bürgin Schaller, R.; Farkas, M.; Kocsis, T.; Clerc, J. T.; Pretsch, E.; Aeschimann, R.; Badertscher, M.; Brodmeier, T.; Fürst, A.; Hediger, H.-J.; Junghans, M.; Kubinyi, H.; Munk, M. E.; Schriber, H.; Wegmann, D. *SpecTool: A Hypermedia Book for Structure Elucidation of Organic Compounds with Spectroscopic Methods*; Chemical Concepts: D-69442 Weinheim, Germany, 1994.

(19) Fürst, A.; Clerc, J. T.; Pretsch, E. A computer program for the computation of the molecular formula. *Chemom. Intell. Lab. Syst.* **1989**, *5*, 329–334.

(20) Derendyaev, B. G.; Nekhoroshev, S. A.; Lebedev, K. S.; Kirshansky, 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.

(21) Heuerding, S.; Clerc, J. T. Simple tools for the computer-aided interpretation of mass spectra. *Chemom. Intell. Lab. Syst.* **1993**, *20*, 57–69.

CI980171B