General Characteristics of Good-List and Bad-List Entries for Structure Generators from Spectra

Hans Schriber and Ernö Pretsch*

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

Received February 27, 1997[⊗]

The consequences of incomplete spectra interpretation and the dependence of the efficiency of demanded and forbidden fragments on their structural characteristics are investigated. It is shown that the performance of a rule-based system strongly depends on the type of structure generator applied.

INTRODUCTION

Since the early days of computer-aided structure elucidation, the following three-step process has been devised:[1] (1) generation of the molecular formula and derivation of structural information from spectra, (2) from these, creation of all possible isomers, and (3) estimation of the spectra for the structures thus obtained and comparison with the experimental ones. In ideal cases, module 1 should provide sufficient information to reduce the solution space to less than about 10 000 entries so that step 3, for which mainly ^{13}C-NMR[2−6] and ^{1}H-NMR[7,8] spectra estimations have been used so far, efficiently produces a manageable number of candidates. Module 2, the isomer generator, is intimately integrated into the interpretation part in some cases and fully modular in others. The numerous implementations differ mainly in their capability and efficiency of coping with various kinds of structural information. The most time-consuming part of the algorithms consists of avoiding or detecting the multiple occurrence of structures (isomorphism test). Its efficiency can be greatly enhanced by using an atomic representation of nonoverlapping substructures (macroatoms),[9,10] but the price to be paid is a dramatic loss of performance if other kinds of structural information are applied (potential overlaps, alternative fragments, substructures that must be absent, i.e., bad-list items, etc.). One of the most versatile programs of this kind efficiently dealing with possible partial overlaps is ASSEMBLE.[11] However, programs generating structures by successive bond formation between substructures are not efficient enough to cope with a large number of overlapping or alternative substructures, as, e.g., obtained from 2D NMR spectra. In such cases, a structure reduction approach in which forbidden bonds are successively eliminated from the hyperstructure representation has proved to be much more efficient.[12]

Module 1, which automatically derives substructures from spectral information, is an integral part of various applications. CHEMICS[13] and SESAMI[14] choose the relevant building blocks of their systems according to NMR chemical shifts. The DARC system[15] uses substructure−subspectra correlations to retrieve good-list fragments from ^{13}C-NMR spectral data. STREC/X-PERT,[16] ESSESA,[17] and EXSPEC,[18] on the other hand, apply rule-based systems. Some programs work with fuzzy[19] or ambiguous[20] information. Several systems apply more than one spectroscopic method in their rule base[13,14,16,21,22] or in a sense by using C−H COSY data[13,2] but, in general, only one after the other, although the power of simultaneous use has been demonstrated recently.[24] Besides routines as integral parts of structure generators, some modular systems automatically derive substructures that must be present (good-list items) or absent (bad-list items) as, for example, PAIRS[25] and EXSPEC,[18] which use IR spectroscopy, and MSClass[26] employing mass spectrometry.

The fundamental problem of automatic interpretation systems is that a 100% reliability is demanded because one single erroneously predicted bad-list or good-list item leads to the fatal error of missing the correct structure. Therefore, interpretations must be highly conservative, which, of course, reduces the power of the method. Although a certain substructure is either present in, or absent from, the target molecule, it is advisable that interpretation systems should not impose definite statements, i.e., they should allow an answer to be rejected in doubtful cases.[26] Systems relying on databases of measured spectra are not comprehensive for obvious reasons. This does not necessarily lead to a loss of valid structures, if the database is used for substructure identification (module 1).[27−29] However, 100% reliability is not granted if the whole generation process relies on a database, although such systems can be impressively efficient in favorable cases[30] if signal assignments are updated during the structure generation.[15,30]

In this work, the reliability and efficiency of fragments as good-list or bad-list items are investigated with a view to designing a rule-based system. Taking into account the results obtained, a set of 82 rules has been developed to predict the presence or absence of substructures by simultaneously interpreting ^{1}H-NMR, ^{13}C-NMR, and IR spectra.[31] Within the scope of the reference databases employed, a 100% reliability was achieved. A set of examples showed that the capability of the rules to reduce the solution space depends on the type of structure generator used in module 2.

RESULTS

Types of Errors. The basic problem of automatically generating substructures that must be present or absent is that substructure−subspectra correlations are based on current knowledge and that their application to an individual case may be outside their scope; i.e., the use of a rule-based system potentially implies an extrapolation. Therefore, even

[⊗] Abstract published in *Advance ACS Abstracts,* August 15, 1997.

S0095-2338(97)00013-9 CCC: $14.00

Table 1. Various Types of Substructure−Subspectra Correlations and the Consequences of Incomplete Interpretations[a]

	spectral feature	corresponding fragment	conclusion	consequences
1	**X**	**A**	X implies A (good-list entry)	OK
			$\bar{X}$ implies $\bar{A}$ (bad-list entry)	OK
2	**X**	**A** or **B**	X implies A (good-list entry)	possibly fatal error
			$\bar{X}$ implies $\bar{A}$ (bad-list entry)	OK (but rule not effective because no statement about B)
3	**X** or **Y**	**A**	X implies A (good-list entry)	OK (but rule not effective since no statement about Y)
			$\bar{X}$ implies $\bar{A}$ (bad-list entry)	possibly fatal error
4	**X** or **Y**	**A** or **B**	X implies A (good-list entry)	possibly fatal error
			$\bar{X}$ implies $\bar{A}$ (bad-list entry)	possibly fatal error

[a] $\bar{X}$, $\bar{A}$ mean that X, A are missing.

if such an interpretation system is designed with extreme precaution, a 100% reliability can never be guaranteed. Various types of incomplete interpretations are, however, conceivable, and not all result in fatal errors, as is shown in the following.

Four distinct possibilities are presented in Table 1. Case 1 represents the ideal situation; i.e., the spectral feature, X, such as an IR absorption or an NMR chemical shift within certain spectral ranges or a given combination of fragment ions in MS, corresponds uniquely to the structural element, A. This means that A *uniquely* and *always* induces the feature X. In this case, and only then, the presence or absence of X always implies the presence or absence of A, respectively. Some programs described in the literature assumed this ideal case to hold throughout; i.e., if A is not found as a good-list entry, it is automatically used as a bad-list item.[17] However, since such ideal cases are seldom encountered, it is advisable to allow the statement "neither good list nor bad list" as a valid answer of the system to avoid an error if the conclusion is ambiguous.[26] If two different fragments, A and B, may induce a spectral feature X and they always do so, the absence of X is correctly interpreted by including A in the bad list (Table 1, case 2). Although this interpretation is correct *per se*, it is incomplete because no statement is made about the absence of B. In such a case, to interpret the occurrence of the spectral feature X as evidence for the presence of the structural feature A (or B) may be a fatal error. For example, it is known that cyclic iminoethers induce a strong IR absorption around 1680 cm^{-1}, not to be distinguished from a C═O band even by experienced spectroscopists. Therefore, to interpret the presence of such an absorption as a C═O indicator may lead to a fatal error, whereas the absence of the band within the appropriate range is correctly interpreted as the absence of carbonyl. However, the efficiency of this bad-list rule is not optimal if no statement has been included about iminoethers. Case 3 in Table 1 corresponds to the counterpart of case 2 and is mainly of academic interest. The substructure A induces either the spectral feature X or Y, and neither of them occur if A is absent. Hence, to infer the presence of A solely on the basis of X, although correct, is incomplete and, therefore, not of optimal efficiency. On the other hand, to assume the absence of A if X is absent may again produce a fatal error. Case 4 shown in Table 1 unfortunately is most common in spectroscopy: One spectral feature may be caused by several fragments, each of which may induce different spectral features so that any prediction about the presence or absence of a fragment is risky.

Efficiency. So far, no study is known of the relationship between the characteristics of substructures and their power to reduce the solution space when used as good-list or bad-list items. Here, as a quantitative measure of efficiency, the reduction factor for any molecular formula given is defined as the ratio (in percent) between the number of structures that remain after applying one single bad-list or good-list rule, and the total number of possible structures:

$$\text{reduction factor} = 100\,\frac{\text{number of structures left}}{\text{total number of possible structures}} \quad (1)$$

Although the reducing power of a substructure as a good-list or bad-list item depends on the specific problem, i.e., on the occurrence of elements in the molecular formula, of double bond equivalents in the molecule, etc., some general rules could be derived from the examples shown in Figure 1. Obviously, the same fragment used as a good-list or a bad-list entry leads to complementary results; however, its power to reduce the solution space greatly varies in both cases. For example, by requiring the presence of a methylene group (first entry), the total number of 607 376 possible structures is only lowered to 607 274, whereas its absence causes a reduction to 102. The opposite holds if the fragment is a monosubstituted phenyl group (last entry), which is present in only 148 of the possible structures. Figure 1 indicates as a general trend that small fragments with many free valences are more efficient as bad-list entries, while large ones with few free valences are especially useful as good-list entries.

The influence of free valences on the efficiency of individual fragments is evident from the results shown in Figure 2 for the molecular formula C_9H_{12}, yielding 19 983 different structures. If the saturated C_2 fragments have four or five free valences (Figure 2, items 1−3), they are very efficient as bad-list items but nearly useless in the good-list. By successively reduction of the number of their free valences, the fragments improve their capability as good-list entries, while their usefulness in the bad list deteriorates because they occur in a smaller number of structures.

The type of elements in the fragments also influences their power to reduce the solution space. The results shown in Figure 3 were obtained for the molecular formula $X_4Y_3Z_2H_{18}$, yielding 255 774 possible structures with one double bond equivalent, the three hypothetical elements (X, Y, and Z) being tetravalent in order to avoid any bias by different valences. In all five sets of fragments, the same trend is observed:

A good-list fragment containing a less frequent element (Z in Figure 3) is more efficient than another one with an element of higher occurrence, the number of the structure left for the above molecular formula decreasing in the sequence of X → Y → Z. The opposite, of course, holds for these elements as part of bad-list fragments. Therefore, heteroatoms, which in organic molecules usually occur less frequently than C atoms, are more efficient in bad-list than in good-list items if all other factors remain unchanged. From

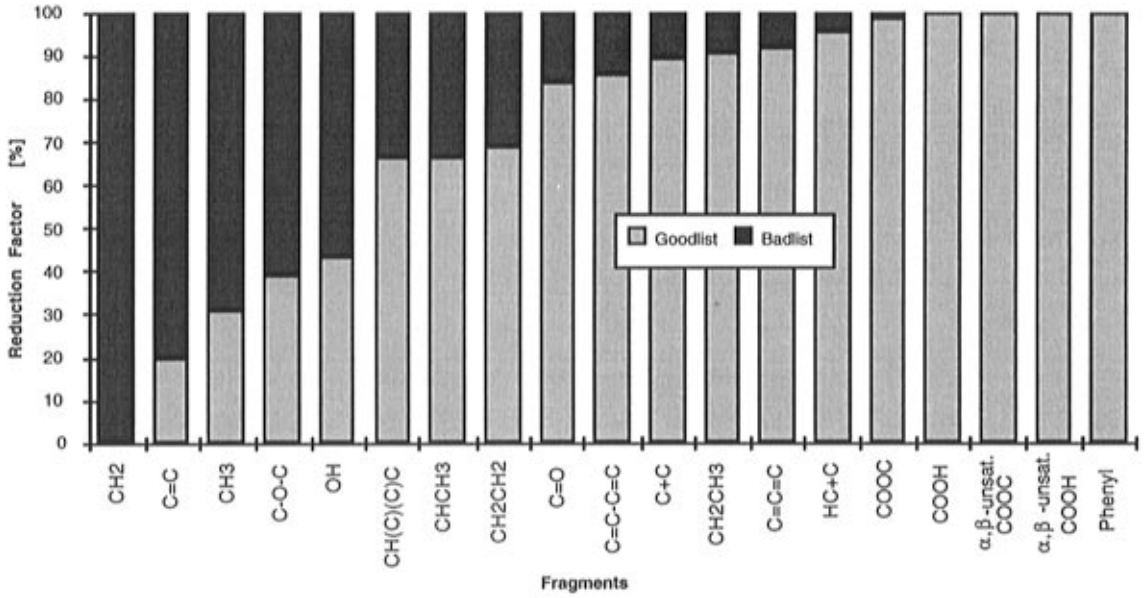
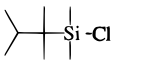
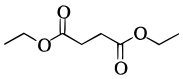
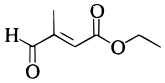
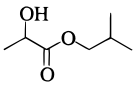
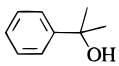
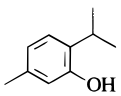


Figure 1. Influence of fragments, individually used as good-list or bad-list item on the reduction factor calculated with the programs MOLGEN[32] and ASSEMBLE[33] for the molecular formula $C_8H_{10}O_2$. Total number of possible structures: 607 376.

Table 2. Reduction in the Total Number of Possible Structures (Eq 1) by Various Structure Generators Using Structural Information Provided by the Rule-Based System Described

molecular formula of unknown	total no. of possible structures	no. of structures left (reduction factor, %) using the structure generator				correct structure
		MOLGEN[32]		ASSEMBLE[11,33]		
		version 3.0	version 3.1	incl[a]	excl[a]	
$C_8H_{19}ClSi$	1 608	265 (83.5)	26 (98.4)	6 (99.6)	6 (99.6)	
$C_8H_{14}O_4$	2 224 538	47 419 (97.9)	4 724 (99.8)	41 (99.998)	39 (99.998)	
$C_4H_8N_2O$	6 754	240 (96.4)	115 (98.3)	103 (98.5)	101 (98.5)	
$C_5H_4O_2$	1 821	311 (82.9)	97 (94.7)	22 (98.8)	7 (99.6)	
$C_5H_4O_3$	7 744	281 (96.4)	281 (96.4)	373 (95.1)	118 (98.5)	
C_6H_6OS	28 521	2 120 (92.6)	893 (96.9)	857 (97.0)	377 (98.7)	
$C_7H_{10}O_3$	308 660	10 271 (96.7)	1 900 (99.4)	1 267 (99.6)	995 (99.7)	
$C_7H_{14}O_3$	22 151	5 087 (77.0)	1 449 (93.5)	38 (99.8)	38 (99.8)	
$C_9H_{12}O$	338 761	38 892 (88.5)	1 675 (99.5)	338 (99.9)	133 (99.96)	
$C_{10}H_{14}O$	1 548 361	32 371 (97.9)	2 564 (99.8)	212 (99.99)	85 (99.99)	

[a] Including or excluding improbable structures.

the examples in Figure 3, the same influence of free valences is observed as stated above. In addition, it is seen that unsaturated fragments are more efficient as good-list items (in parallel to the trend expected owing to the lower number of free valences) than the corresponding saturated ones.

The chemical environment of the substructures, such as the number of neighboring H atoms, can often be derived from spectroscopic information, but most structure generators developed so far are unable to make use of it. It could be defined by logical OR combinations of a set of substructures, but this procedure would be rather cumbersome and error-prone. Fortunately, some programs, such as ASSEMBLE,[11] support the definition of the surroundings of a fragment used as a good-list item. A well-defined neighborhood logically

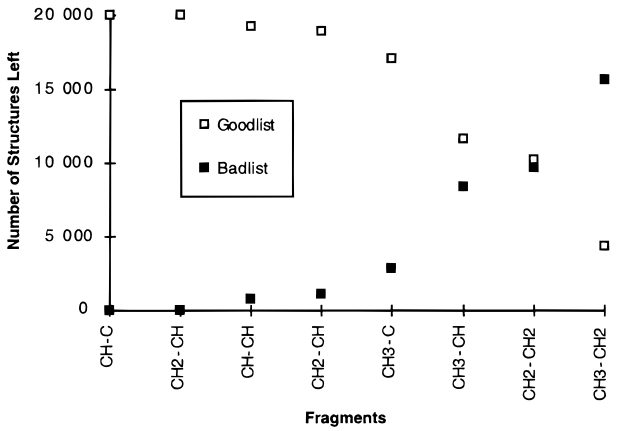


Figure 2. Influence of free valences of a series of fragments on the number of structures left after using them individually as good-list or bad-list items (molecular formula: C_9H_{12}, total number of possible structures: 19 983).

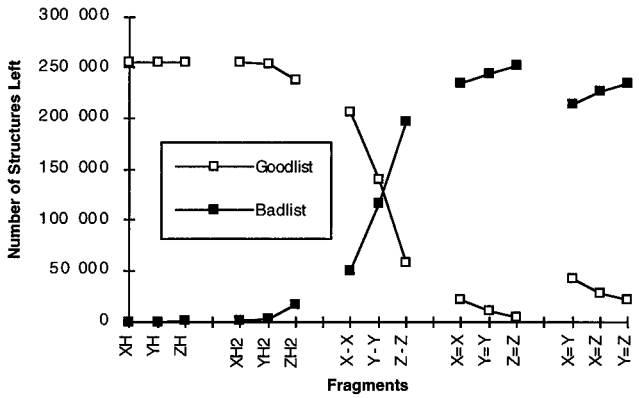


Figure 3. Influence of the type of atoms occurring in substructures on their efficiency as good-list or bad-list items using $X_4Y_3Z_2H_{18}$ as hypothetical molecular formula (X, Y, and Z: different kinds of tetravalent atoms).

implies a larger substructure which thus, in comparison to one with undefined neighbors, is expected to lead to a good-list item of higher efficiency. The molecular formula $C_8H_{18}O$, e.g., corresponds to 171 structures if chemically improbable ones are excluded.[11] All but 6 of them have a CH_3CH group so that this group would be of little use as a good-list entry (note that CH_3CH may occur as an embedded group of CH_3CH_2). However, by further demanding that its neighboring atoms are not attached to H atoms, only 2 structures are left. Thus, the efficiency of this group as a good-list item is extremely enhanced.

Influence of Structure Generators. On the basis of the above-mentioned results, a set of 82 rules have been derived that infer the presence or absence of a total of 60 substructures whose neighborhood is taken into account as well.[31] Every rule was checked against available databases, and no violation was observed. The efficiency of the rules, however, depends on how far their results can be used by the structure generator. This is illustrated by Table 2, which gives the results obtained with different generators by applying the rules on the spectra of 21 compounds. In the first example, $C_8H_{19}ClSi$, an isopropyl group occurs whose methyl groups are the only vicinal coupling partners of the methine proton. The molecule has an additional methyl and a quaternary C atom, but no ethyl group. Its ^{13}C-NMR spectrum reveals a symmetry leading to 5 signals. From all this, version 3.0 of MOLGEN[32] only used the information about the presence of an isopropyl and the absence of an ethyl group and generated 265 out of the 1608 structures possible. With version 3.1, the reduction factor was substantially increased because the information about the number of CH_3, CH_2, and CH groups and of C atoms (minimums 3, 0, 1, and 1, respectively) was also considered. ASSEMBLE, on the other hand, capable of applying the full information listed above, further reduced the number of possible structures from 26 to 6. For the second example, $C_8H_{14}O_4$, the information generated by the rules was the presence of an ethyl group without further coupling partners, of a methylene group, and of a quaternary C atom. The bad-list items were OH, a triple bond between two C atoms, a terminal vinyl, a methoxy, and an aldehyde group. Finally, a symmetry yielding a total of four anisochronous C atoms was demanded. While ASSEMBLE made use of the full information leading to 41 of the 2 224 538 possible structures, versions 3.0 and 3.1 of MOLGEN considered only parts of it and generated 47 419 and 4724 structures, respectively. Similar tendencies were observed in all 21 examples investigated (only 10 of them shown) when the two structure generators in their different modifications were applied. The reduction factors obtained with versions 3.0 and 3.1 of MOLGEN and with ASSEMBLE (including improbable structures) are 90.8, 97.5, and 98.9% (means) or 93.4, 98.4, and 99.4% (medians), respectively. The exclusion of chemically improbable solutions with ASSEMBLE gave a slight improvement of 0.3 (mean) and 0.2% (median). These data show the importance of applying structure generators that make full use of the complete spectroscopic information available.

ACKNOWLEDGMENT

We thank Prof. A. Kerber and R. Laue for providing us with different versions of MOLGEN and Prof. Dr. M. E. Munk for ASSEMBLE. This work was partly supported by the Swiss National Foundation. Thanks are also due to Dr. D. Wegmann for careful reading of the manuscript.

REFERENCES AND NOTES

(1) Gray, N. A. B.; Carhart, R. E.; Lavanchy, A.; Smith, D. H.; Varkony, T.; Buchanan, B. G.; White, W. C.; Creary, L. Computerized mass spectrum prediction and ranking. *Anal. Chem.* **1980**, *52*, 1095−1102.
(2) Crandell, C. W.; Gray, N. A. B.; Smith, D. H. Structure evaluation using predicted ^{13}C spectra. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 48−57.
(3) Dubois, J. E.; Sobel, Y. DARC System for documentation and artificial intelligence in chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 326−333.
(4) Funatsu, K.; Miyabayashi, N.; Sasaki, S. Further Development of the Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18−28.
(5) Neudert, R.; Penk, M. Enhanced Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 244−248.
(6) Pretsch, E.; Fürst, A.; Badertscher, M.; Bürgin, R.; Munk, M. E. C13Shift: A computer program for the prediction of ^{13}C-NMR spectra based on an open set of additivity rules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 291−295.
(7) Funatsu, K.; Acharya, B. P.; Sasaki, S. Application of a digital ^{1}H-NMR spectrum to the survival test of substructures and the assignment test. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 735−744.
(8) Bürgin Schaller, R.; Munk, M. E.; Pretsch, E. Spectra estimation for computer-aided structure determination. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 239−243.
(9) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Application of artificial intelligence for chemical inference XVII. An approach to computer-assisted elucidation of molecular structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755−5762.
(10) Wieland, T.; Kerber, A.; Laue, R. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413−419.

(11) Shelley, C. A.; Munk, M. E. CASE, a computer model of the structure elucidation process. *Anal. Chim. Acta* **1981**, *133*, 507−516.
(12) Christie, B. D.; Munk, M. E. Structure generation by reduction: A new strategy for computer-assisted structure elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87−93.
(13) Funatsu, K.; Sasaki, S. Recent advances in the automated structure elucidation system CHEMICS. Utilization of two-dimensional NMR spectral information and development of peripheral functions for examination of candidates. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.
(14) Munk, M. E.; Velu, V. K.; Madison, M. S.; Robb, E. W.; Badertscher, M.; Christie, B. D.; Razinger, M. In *Recent Advances in Chemical Information*; Collier, H., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1993; Vol. II.
(15) Dubois, J. E.; Carabedian, M.; Dagane, I. Computer-aided elucidation of structures by carbon-13 NMR. *Anal. Chim. Acta* **1984**, *158*, 217−233.
(16) Elyashberg, M. E.; Martirosian, E. 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.
(17) Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra. 6. Substructure constraints from analysis of ^{13}C-NMR Spectra. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 979−1000.
(18) Luinge, H. J.; Kleywegt, G. J.; van't Klooster, H. A.; van der Maas, J. H. Artificial intelligence used for the interpretation of combined spectral data. 3. Automated generation of interpretation rules for infrared spectral data. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95−99.
(19) Laidboeur, T.; Laude, I.; Cabrol-Bass, D.; Bangov, I. P. Employment of fuzzy information derived from spectroscopic data toward reducing the redundancy in the process of structure generation. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 171−178.
(20) Peng, C.; Yuan, S.; Zheng, C.; Shi, Z.; Wu, H. Practical computer-assisted structure elucidation for complex natural products: Efficient use of ambiguous 2D NMR correlation information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 539−546.
(21) Gray, N. A. B. Structural interpretation of spectra. *Anal. Chem.* **1975**, *47*, 2426−2431.
(22) Luinge, H.-J.; Maas, J. H. v. d. Artificial intelligence for the interpretation of combined spectral data. *Anal. Chim. Acta* **1989**, *223*, 135−147.
(23) Christie, B. D.; Munk, M. E. The role of two-dimensional nuclear magnetic resonance spectroscopy in computer-enhanced structure elucidation. *J. Am. Chem. Soc.* **1991**, *113*, 3750−3757.
(24) Munk, M. E.; Madison, M. S.; Robb, E. W. The neural network as a tool for multispectral interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.
(25) Woodruff, H. B.; Smith, G. M. Computer program for the analysis of infrared spectra. *Anal. Chem.* **1980**, *52*, 2321−2327.
(26) Varmuza, K.; Werther, W. Mass spectral qualifiers for supporting systematic structure elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323−333.
(27) Shelley, C. A.; Munk, M. E. Computer prediction of substructures from carbon-13 nuclear magnetic resonance spectra. *Anal. Chem.* **1982**, *54*, 516−521.
(28) Bremser, W.; Fachinger, W. Multidimensional spectroscopy. *Magn. Res. Chem.* **1985**, *23*, 1056−1071.
(29) Robien, W. Computer-assisted structure elucidation of organic compounds III. Automatic fragment generation from 13C-NMR spectra. *Mikrochim. Acta II* **1986**, 271−279.
(30) Will, M.; Fachinger, W.; Richert, J. R. Fully automated structure elucidation-A spectroscopist's dream comes true. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.
(31) Schriber, H.; Pretsch, E. Rule-based system to derive automatically good-list and bad-list entries for structure generators from spectra. *J. Chem. Inf. Comput. Sci.*, following paper in this issue.
(32) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Anal. Chim. Acta* **1992**, *314*, 141−147.
(33) Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. An approach to automated partial structure expansion. *Anal. Chim. Acta* **1978**, *103*, 121−132.

CI9700135