Spectra Estimation for Computer-Aided Structure Determination[†]

Renate Bürgin Schaller,[‡] Morton E. Munk,[§] and Ernö Pretsch*[,‡]

Department of Organic Chemistry, Swiss Federal Institute of Technology (ETH),
Universitätstrasse 16, CH-8092 Zürich, Switzerland, and Department of Chemistry, Arizona State University,
Tempe, Arizona 85287

Received September 30, 1995[⊗]

A recently developed program for estimating ^{1}H-NMR chemical shifts has been interfaced to a structure generator. It provided predicted chemical shifts for 89% of the protons occurring in ca. 110 000 different chemical environments of 24 308 generated structures. It is possible to rank the structures by comparing measured and estimated chemical shifts.

INTRODUCTION

Depending on the particular task, two different schemes of the interpretation of spectra are conceivable. *Structure verification* is possible if both the proposed structure and its spectra are available (Figure 1, left). It is implemented by structure search in a spectroscopic library or, if the compound is not registered, by spectra estimation. Retrieved or estimated spectra are then compared with the measured one. In absence of a proposed structure, *structure elucidation* can be performed on the basis of spectra (Figure 1, right). A spectrum search in spectroscopic databases may solve the problem if the unknown compound is documented. Otherwise, pieces of structural information have to be collected and an exhaustive list of possible isomers must be generated which can then be ranked on the basis of spectrum estimation followed by a comparison with the measured spectrum.

The different boxes in Figure 1 represent computational tools for these processes. Numerous spectroscopic databases are available today for routine usage. During the last two decades, over 20 isomer generators have been described.[1−3] Some of them are ready for routine use although their efficiency and flexibility concerning the information to be entered should still be improved. In spite of numerous efforts and some exaggerated claims in the literature, the automatic derivation of structural elements from spectra[4] remains the least successful part of computer-aided structure elucidation. Interpretation by experts is still superior in all cases investigated.[5] For an automatic structure elucidation system, absolute reliability is a prerequisite since one single wrong statement about the presence or absence of a substructure would be fatal.

Tools for fast and reliable estimations of ^{1}H-, ^{13}C-NMR, mass, and IR spectra are required both for structure verification and elucidation (cf. Figure 1). The most important characteristics of such programs should be reliability and broad scope. Especially if structure generators are used, spectrum estimations for unusual structures are needed. All spectrum estimation methods, including the most sophisticated quantum chemical methods, are based on some empirical knowledge. Spectrum estimation of uncommon structures might present problems if the empirical rules applied are not valid for the target structure. Therefore, reliable programs should signal their limits and refuse to estimate rather than give uncertain results.

In our group, programs based on simple linear models have been developed for predicting ^{1}H- and ^{13}C-NMR spectra.[6−8] Entirely different strategies are followed in current projects leading to the prediction of IR and mass spectra. Here, we report on the combination of the Proton Shift program with the structure generator ASSEMBLE[9] which has been realized within SpecTool.[10]

PROTON SHIFT

The recently developed program, Proton Shift, for predicting ^{1}H-NMR chemical shifts is based on simple linear models. It automatically detects the substructures for which rules are available and identifies, for each of them, the remaining part of the structure as a set of substituents. Various extrapolation steps, including the disassembly of large substructures and replacement of missing substituents by embedded ones or by assigning substituents of related substructures, were applied to improve the scope of the program. In addition, a set of new increments have been developed.[11] The program currently relies on some 3000 parameters.

For a limited test set of 583 chemical shifts, the mean deviation between predicted and measured values was −0.06 ppm, with a standard deviation of 0.18 ppm. Since no independent test set is available, only an estimate of the general quality of prediction can be given. The estimated standard deviation between predicted and measured values is around 0.3 ppm. Except for some cases, such as cis/trans geometry of double bonds and axial/equatorial substituents of six-membered rings, the estimation is based on the constitution (connectivity) only. Therefore, no significant improvements in the precision of prediction can be expected with this model.

An interface has been developed between the structure generator ASSEMBLE and Proton Shift. The constitutional isomers obtained with ASSEMBLE can be imported and ranked in order of decreasing probability of being correct by comparing the estimated shifts with the observed ones. With this new feature, several tests regarding the scope and accuracy of Proton Shift are possible. To investigate its scope, no reference database is necessary. A total of 24 308 isomers were generated with ASSEMBLE (cf. Table 1). The

[†] Dedicated to Professor Shin-Ichi Sasaki on the occasion of his 70th birthday. His ongoing contributions over decades have significantly influenced the field of computer-aided structure elucidation.

[‡] Swiss Federal Institute of Technology (ETH).

[§] Arizona State University.

[⊗] Abstract published in *Advance ACS Abstracts,* February 1, 1996.

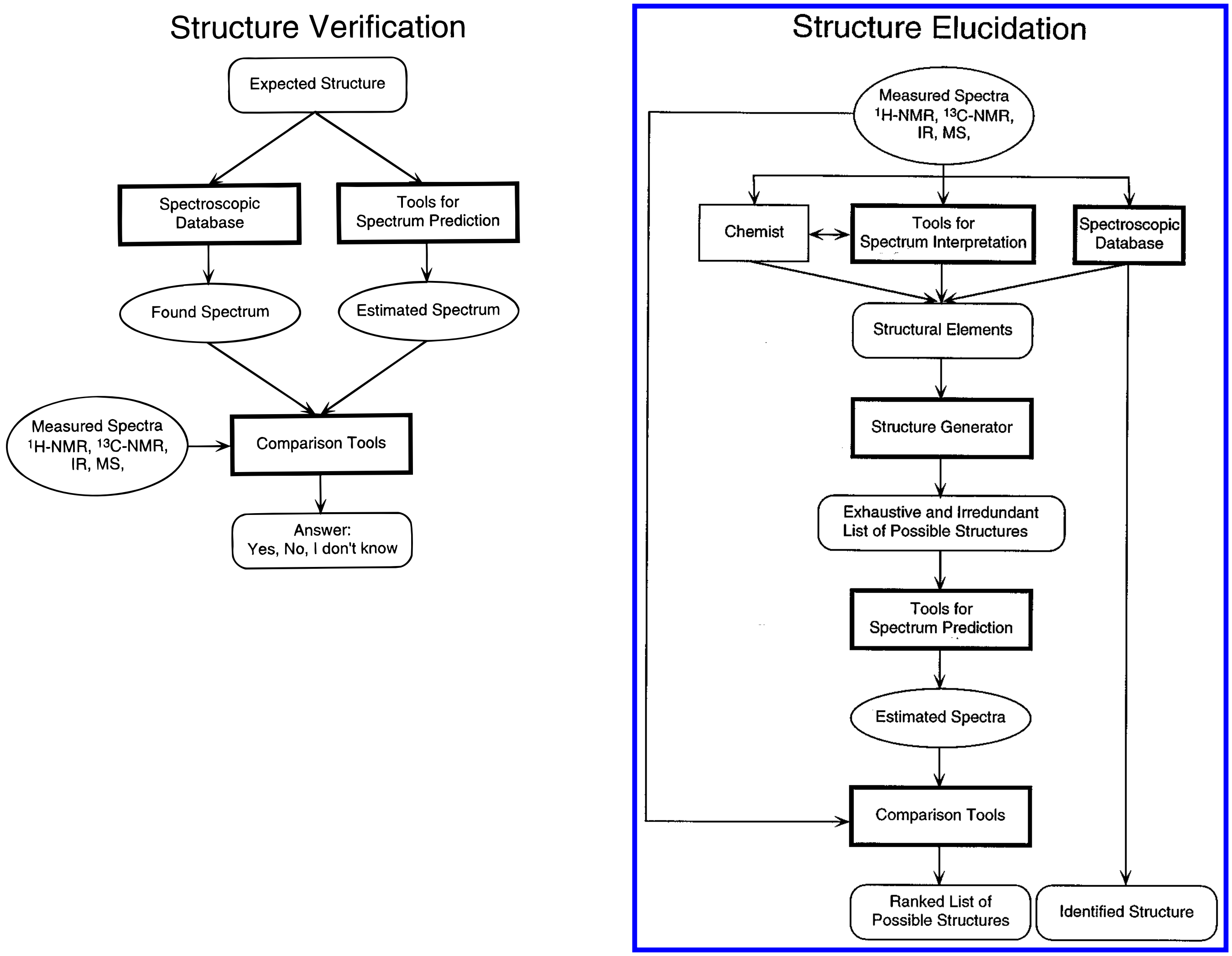

Structure Verification
Expected Structure
Spectroscopic Database
Tools for Spectrum Prediction
Found Spectrum
Estimated Spectrum
Measured Spectra ^{1}H-NMR, ^{13}C-NMR, IR, MS,
Comparison Tools
Answer: Yes, No, I don't know
Structure Elucidation
Measured Spectra ^{1}H-NMR, ^{13}C-NMR, IR, MS,
Chemist
Tools for Spectrum Interpretation
Spectroscopic Database
Structural Elements
Structure Generator
Exhaustive and Irredundant List of Possible Structures
Tools for Spectrum Prediction
Estimated Spectra
Comparison Tools
Ranked List of Possible Structures
Identified Structure

Figure 1. Schematic representation of the structure verification and structure elucidation process.

protons of these compounds occurred in 111 605 different chemical environments. A chemical shift estimation was possible in 89.1% of the cases. Considering the presence of many uncommon structures, this test result is very satisfactory. As indicated in Table 1, the average estimation time on a Macintosh Quadra 650 was 0.30 s per compound. Thus, the program seems to be adequate for practical use in connection with structure generators.

In order to rank the generated structures, their predicted chemical shifts are compared with the measured ones. These are entered manually for the distinct groups of CH_3, CH_2, and CH and are then assigned to the protons of each structure by applying an algorithm that provides the closest possible match. For each of the three CH_n groups, a sorted list of the experimental shifts is created and compared with that of the predicted values. The standard deviations hereby calculated are then used for automatically ranking the structures. This approach evidently does not guarantee accurate shift assignments, but it never discriminates against the correct solution since wrong attributions can only lead to a ranking that is too good but never too bad. In case not all chemical shifts can be predicted, again the assignment with the smallest deviation between measured and estimated values is selected. Structures, for which none of the chemical shifts can be estimated, are ranked above the others. Since the isomer generator applied can only build constitutions, it provides no information about cis/trans positions. In such cases, the

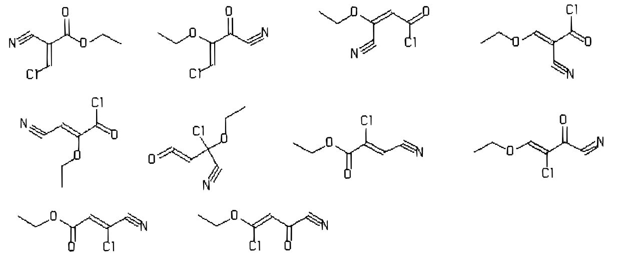


Figure 2. The 10 constitutions (m1−m10) generated for the molecular formula $C_6H_6NO_2Cl$ and some constraints. Arbitrary configurations are drawn for the constitutions generated by ASSEMBLE.

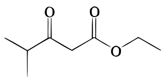


Figure 3. Structure of the compound of molecular formula $C_8H_{14}O_3$ with three CH_3, two CH_2, one CH, and two C=O groups ranked first out of 231 isomers

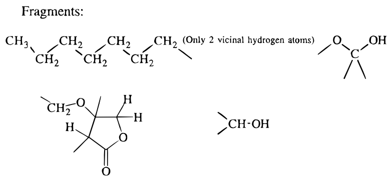


Figure 4. Pieces of information used to generate the 20 isomers shown in Figure 5 having the molecular formula $C_{15}H_{24}O_6$ with the constraints that three rings must be present, two of them five-membered.

Table 1. Scope and CPU Time Needed for the Chemical Shift Estimation of the Different Isomers of Molecular Formulas $C_6H_{4-14}O_{0-2}$ with Proton Shift[a]

molecular formula	no. of isomers	CPU time (s) per isomer[b]	no. of chemical shifts	estimation		
				full	partial	none
C_6H_n						
C_6H_{14}	5	0.60	29	29	0	0
C_6H_{12}	25	0.32	146	146	0	0
C_6H_{10}	68	0.25	375	355	17	3
C_6H_8	77	0.17	408	394	9	5
C_6H_6	47	0.11	204	184	8	12
C_6H_4	5	0.20	17	17	0	0
C_6H_nO						
$C_6H_{14}O$	32	0.47	183	183	0	0
$C_6H_{12}O$	211	0.37	1200	1176	24	0
$C_6H_{10}O$	680	0.32	3616	3313	281	22
C_6H_8O	982	0.29	4865	4444	370	51
C_6H_6O	675	0.19	2844	2481	269	94
C_6H_4O	110	0.10	358	278	68	12
$C_6H_nO_2$						
$C_6H_{14}O_2$	179	0.50	1017	1017	0	0
$C_6H_{12}O_2$	1313	0.62	7290	7067	223	0
$C_6H_{10}O_2$	4527	0.41	23366	21 880	1387	99
$C_6H_8O_2$	7557	0.27	35 623	31 307	3989	327
$C_6H_6O_2$	6122	0.21	24 721	20 969	3175	577
$C_6H_4O_2$	1693	0.20	5343	4273	856	214
sum	24 308		111 605	99 513	10 676	1416
mean		0.30	100.0%	89.1%	9.6%	1.3%

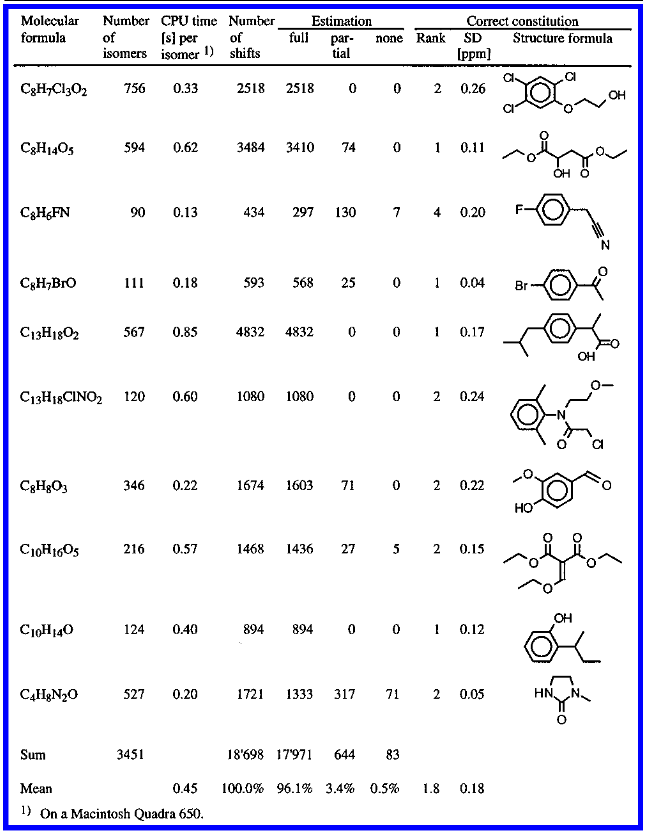
[a] Structure generation was limited to chemically relevant constitutions. Those unstable according to present knowledge were excluded. [b] On a Macintosh Quadra 650.

Table 2. Ranking of the 10 Isomers of Molecular Formula $C_6H_6NO_2Cl$ Shown in Figure 2 with (Bottom) and without (Top) Applying the Procedure That Selects the Configuration Giving the Best Match[b]

	chemical shifts			SD (ppm)
	CH	CH_2	CH_3	
measured	8.10	4.35	1.40	
m4	8.23	4.00	1.22	0.24
m1	7.54	4.19	1.30	0.34
m10	7.39	4.00	1.22	0.47
m6	a	3.41	1.11	0.70
m7	6.76	4.19	1.30	0.78
m9	6.16	4.19	1.30	1.13
m2	6.00	4.00	1.22	1.23
m3	5.84	4.00	1.22	1.32
m5	5.46	4.00	1.22	1.54
m8	5.28	4.00	1.22	1.64
m1	7.89	4.19	1.30	0.16
m4	8.23	4.00	1.22	0.24
m10	7.51	4.00	1.22	0.41
m6	*a*	3.41	1.11	0.70
m7	6.91	4.19	1.30	0.70
m9	6.66	4.19	1.30	0.84
m2	6.03	4.00	1.22	1.22
m3	5.90	4.00	1.22	1.29
m5	5.77	4.00	1.22	1.36
m8	5.37	4.00	1.22	1.59

[a] Chemical shift could not be estimated owing to lack of parameters. [b] Compound m1, shown as best hit in the bottom part, is the correct solution.

Table 3. Scope, CPU Time, and Quality of the Chemical Shift Estimation with Proton Shift for Various Sets of Isomers



Molecular formula	Number of isomers	CPU time [s] per isomer [1)	Number of shifts	Estimation			Correct constitution		
				full	par-tial	none	Rank	SD [ppm]	Structure formula
$C_8H_7Cl_3O_2$	756	0.33	2518	2518	0	0	2	0.26	
$C_8H_{14}O_5$	594	0.62	3484	3410	74	0	1	0.11	
C_8H_6FN	90	0.13	434	297	130	7	4	0.20	
C_8H_7BrO	111	0.18	593	568	25	0	1	0.04	
$C_{13}H_{18}O_2$	567	0.85	4832	4832	0	0	1	0.17	
$C_{13}H_{18}ClNO_2$	120	0.60	1080	1080	0	0	2	0.24	
$C_8H_8O_3$	346	0.22	1674	1603	71	0	2	0.22	
$C_{10}H_{16}O_5$	216	0.57	1468	1436	27	5	2	0.15	
$C_{10}H_{14}O$	124	0.40	894	894	0	0	1	0.12	
$C_4H_8N_2O$	527	0.20	1721	1333	317	71	2	0.05	
Sum	3451		18'698	17'971	644	83			
Mean		0.45	100.0%	96.1%	3.4%	0.5%	1.8	0.18	

[1) On a Macintosh Quadra 650.

chemical shifts of all possible stereoisomers are calculated, and again the match closest to the experimental value is used for ranking. The importance of this procedure is documented by Figure 2 and Table 2. The 10 isomers generated for $C_6H_6NO_2Cl$ by entering the molecular formula and some substructure information are shown in Figure 2. The different configurations of the double bond have to be taken into account since ASSEMBLE only generates constitutions. By using the procedure outlined above, the correct constitution is ranked best, with a standard deviation of 0.16 ppm between observed and measured shifts (Table 2, bottom). If the procedure is not applied, the correct structure is ranked in second position, with a much larger deviation of 0.34 ppm (Table 2, top).

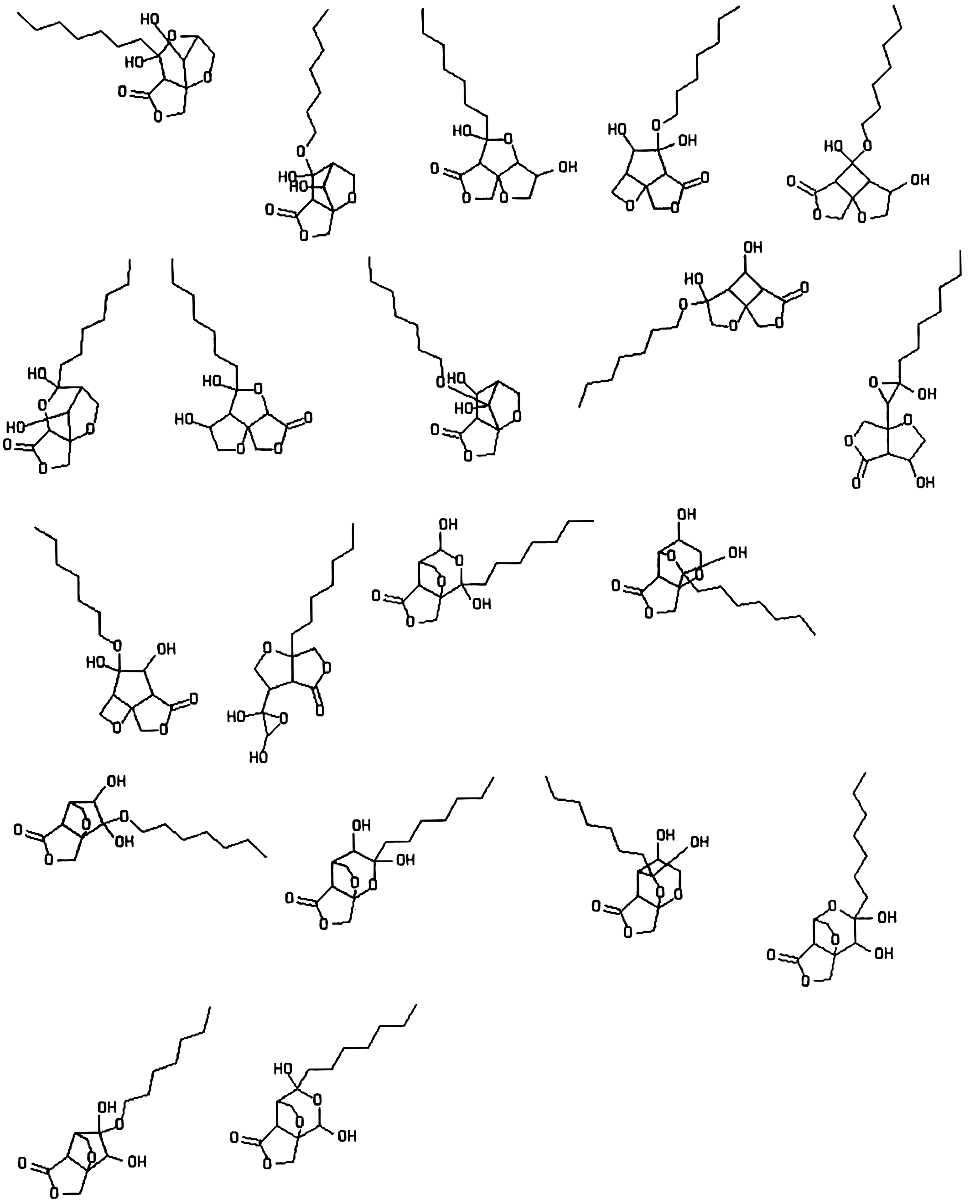


Figure 5. The 20 constitutions generated using the information shown in Figure 4. The correct compound is the second entry in row 2.

The accuracy of the estimation program Proton Shift and the effectiveness of the procedure of assignment were tested on several sets of examples. First, all 231 chemically reasonable isomers with three CH_3, two CH_2, one CH, and two C=O groups were generated for the molecular formula $C_8H_{14}O_3$. Based on the observed chemical shift values of 1.15, 1.15, 1.30 (methyl), 3.50, 4.20 (methylene), and 2.70 (methine) entered, the correct solution shown in Figure 3 was ranked first, with a standard deviation of 0.05 ppm.

In Table 3, a series of test cases are given. For each of the compounds, a large number of isomers was generated (between 90 and 756). Of 18 698 different chemical environments, 96.1% of the chemical shifts could be estimated. The correct structures were ranked in positions 1 (four cases), 2 (five cases), or 4 (one case) with standard deviations between measured and estimated shifts of 0.04–0.26 ppm (mean overall deviation: 0.18 ppm).

A final example shows that the method is also useful for larger molecules. Based on the information given in Figure 4, ASSEMBLE generated 20 constitutions of molecular formula $C_{15}H_{24}O_6$. The correct solution (second entry in row 2 of Figure 5) was ranked in position 4 with a standard deviation of 0.33 ppm (SD for the first, tenth and last position: 0.20, 0.55, and 0.73 ppm). In this case, differences between the measured chemical shifts of diastereotopic protons (not distinguished by Proton Shift) were up to 0.35 ppm.

CONCLUSIONS

Proton Shift, a recently developed program for estimating ^{1}H-NMR chemical shifts, has been interfaced to the structure generator ASSEMBLE. It is capable of ranking generated structures according to the predicted chemical shifts by comparing them with the experimental values entered. Based on the results, part of the solutions can be safely eliminated. Current work focusses on the prediction of IR and mass spectra. They, together with Proton Shift and the previously developed C13Shift programs, will then allow to rank generated structures on the basis of all four spectroscopic methods.

ACKNOWLEDGMENT

This work was partly supported by the Swiss National Science Foundation. The authors thank Dr. D. Wegmann for careful reading of the manuscript.

REFERENCES AND NOTES

(1) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986.

(2) 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.

(3) Funatsu, K.; Nishizaki, M.; Sasaki, S. Introduction of NOE data to an automated structure elucidation system CHEMICS. Three-dimensional structure elucidation using the distance geometry method. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 745−751.

(4) Funatsu, K.; Acharya, B. P.; Sasaki, S. Application of a digital 1H-NMR spectrum to the survival test of substructures and the assignment test. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 735−744.

(5) Visser, T.; Luinge, H. J.; van der Maas, J. H. Recognition of visual characteristics of infrared spectra by artificial neural networks and partial least squares regression. *Anal. Chim. Acta* **1994**, *296*, 141−154.

(6) Bürgin Schaller, R.; Pretsch, E. A computer program for the automatic estimation of ^{1}H-NMR chemical shifts. *Anal. Chim. Acta* **1994**, *290*, 295−302.

(7) 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.

(8) 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.

(9) Shelley, C. A.; Munk, M. E. CASE, a computer model of the structure elucidation process. *Anal. Chim. Acta* **1981**, *133*, 507−516.

(10) 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, M.; 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, 1994.

(11) Bürgin Schaller, R.; Arnold, C.; Pretsch, E. New parameters for predicting ^{1}H NMR chemical shifts of protons attached to carbon atoms. *Anal. Chim. Acta* **1995**, *312*, 95−105.

CI950141Y