

# Rule-Based System To Derive Automatically Good-List and Bad-List Entries for Structure Generators from Spectra

Hans Schriber and Ernő Pretsch\*

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

Received February 27, 1997<sup>®</sup>

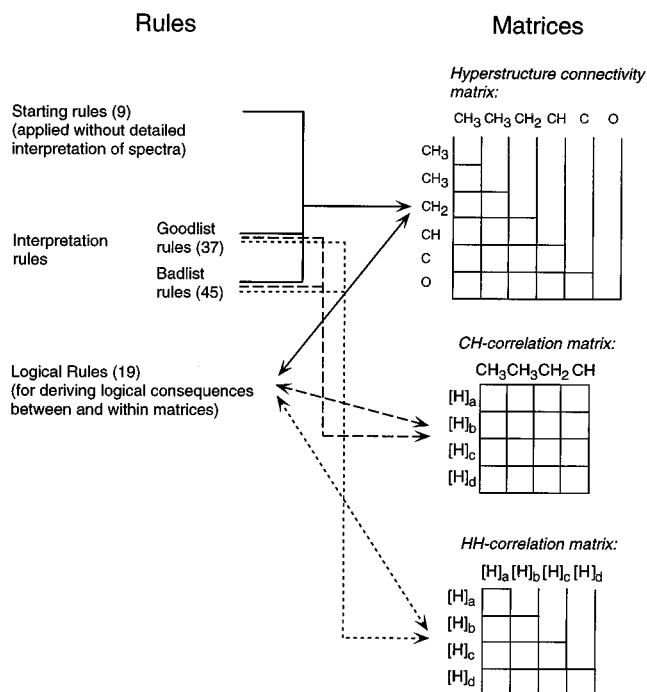
A new rule-based spectra interpretation system named SpecInt is described. Its internal information representation and the simultaneous use of several spectroscopic methods allow one to achieve both high reliability and good performance. For small- and medium-sized molecules, SpecInt is capable of automatically reducing the solution space to under 3%.

## INTRODUCTION

Because of the ever increasing number of new synthetic compounds, the introduction of an automated structure elucidation system is more and more urgent. In most existing approaches, a three-step procedure is conceived by which first the molecular formula and (sub)structural information are generated upon which all compatible isomers are produced. In the third step, the possible solutions thus found are ranked according to the similarity of their predicted and the measured spectra. Today, adequate structure generators are available, and, as a consequence of recent progress in spectra estimation, the automatic ranking of some 1000–10 000 solutions is feasible. However, if the molecular formula is the only information available, the number of possible isomers is on the order of  $10^5$ – $10^6$  even for small- and medium-sized molecules. Therefore, various structure generators use interpretation rules as an integrated part or as a front end.<sup>1–8</sup> In addition, a series of rule-based systems have been described for this purpose.<sup>9–13</sup> Since high reliability of the interpretation and good efficiency in reducing the solution space are contradictory requirements, all systems must make compromises in this respect. Based on the general considerations presented in the previous paper,<sup>14</sup> the new rule-based system SpecInt, which produces good-list and bad-list items from various spectra, is described here. Although very conservative interpretation rules are used showing a 100% reliability when applied to large databases, good performance has been achieved by introducing two novel approaches. The internal representation and the inference machine permit synergetic interactions between rules so that information may be obtained even on the presence or absence of fragments that are not explicitly known to the system. Furthermore, the simultaneous use of <sup>1</sup>H-NMR-, <sup>13</sup>C-NMR- (1D and 2D) and IR-spectroscopic data enhances the power and reliability of the system. For the 80 examples investigated, SpecInt was found to reduce the solution space by more than 97%. In none of these cases was the correct structure missed.

## DESCRIPTION OF SPECINT

**Information Representation and Data Processing.** The knowledge base of SpecInt consists of a set of *interpretation*



**Figure 1.** Overview of data processing in SpecInt (in parentheses: number of rules).

*rules* relating spectral information to the presence or absence of a given substructure. Instead of directly giving the results in the output file, these rules operate on the three matrices shown in Figure 1. The hyperstructure connectivity matrix of the basic fragments, which resembles the bonding adjacency matrix (BAM) used by Christie and Munk in COCOA,<sup>15</sup> (for a related approach see ref<sup>16</sup>), keeps track of all possible connections between the non-hydrogen atoms. A basic fragment consists of a non-hydrogen atom with its H atoms. In the first step, the hyperstructure connectivity matrix is generated from the molecular formula and the <sup>13</sup>C-NMR information (e.g., DEPT spectra). It contains as many carbon centers as there are signals in the <sup>13</sup>C-NMR spectrum. For heteroatoms, the assignment of basic fragments may cause problems because of an uncertain distribution of H atoms. In this case, all possible combinations are generated. For example, if the distribution of two H atoms among the heteroatoms N and O is not known, the groups OH, O, NH<sub>2</sub>, and NH are used as basic fragments. Three assignments are allowed for each of the three bond types (single, double,

<sup>®</sup> Abstract published in *Advance ACS Abstracts*, August 15, 1997.

**Table 1.** Three-Digit Code for Representing Bonds between Two Basic Fragments

bond type designation		bond presence designation	
bond type	position in three-digit code from right	presence of bond	attributed value
single	1	yes	2
double	2	possible	1
triple	3	no	0

and triple): They may be present, possible, or absent. Therefore, each hyperstructure matrix entry is a three-digit number, as shown in Table 1. The entry 002, e.g., signifies (from the right) the presence of a single bond and the absence of a double and a triple bond, while the information that a single and a triple bond are possible, but no double bond, is encoded with 101. This notation is compact, easy to read and program, and allows one to formulate fuzzy connectivity information. Two further matrices, the CH- and the HH-correlation matrices (Figure 1), are used by the system. The former assigns the connectivities between the carbon-centered basic fragments with their respective H atoms. If the NMR signals of several H atoms overlap, they are treated here as one group. The HH-correlation matrix keeps track of proton pairs with vicinal coupling using the descriptors "coupling", "no coupling" and "coupling possible". Results of C-H and H-H COSY experiments can be directly entered into these matrices. The good-list and bad-list rules operate simultaneously on all three matrices, as shown in Figure 1.

**Rules.** The above-described matrix representations allow the application of global and local rules. A global good-list rule makes no specific assignments in the connectivity matrix. For example, if a good-list rule claims the presence of a vinylidene group and the basic fragments CH<sub>2</sub> and/or C occur more than once but none of them can be assigned to that vinylidene group, this good-list rule is *global*. A *local good-list rule* would additionally assign one specific CH<sub>2</sub> and/or C to this group. The corresponding *global bad-list rule* would forbid a double bond between all CH<sub>2</sub> and C basic fragments. Finally, a *local bad-list rule* would only forbid a double bond between one specific pair of CH<sub>2</sub> and C. A comparison shows that in the case of good lists, the local rule provides more powerful information because it allows one to find logical consequences by applying the other rules. The opposite holds for bad lists: Here, the information provided by local rules is weaker because it forbids a bond only for one pair of basic fragments and not for all of them. However, local bad-list rules have the advantage that the system can make direct use of their information. Global good-list items, on the other hand, cannot be stored in the matrices, but their presence is recognized and all further rules use the corresponding information so that interactions with other rules are still possible.

SpecInt uses three kinds of rules, namely, starting, interpretative (good list, bad list), and logical rules. The nine *starting rules* are applied at the beginning of the interpretation process when the hyperstructure connectivity matrix is being set up. As an example, one such rule is given in Scheme 1. It generates two bad-list items (N and NH<sub>2</sub>) by comparing the number of H atoms not bonded to carbon with that of the available heteroatoms. Of course, this information would implicitly be found also by a structure generator, but if it is available right from the beginning of the interpretation

**Table 2.** Good-List Fragments with Number of Rules (Totaling 37)

substructure	no. of rules	substructure	no. of rules
CH <sub>2</sub> (VH0202)CH <sub>2</sub> (VH0202)	2	CH <sub>3</sub> (VH0000)	1
CH(H11)(VH0606)(CH <sub>3</sub> )CH <sub>3</sub>	4	CH(H11)(VH0101)(CH1)	2
CH <sub>3</sub> CH <sub>2</sub> (H22)(VH0303)	4	CH <sub>3</sub> CH(H11)(VH0409)	4
CH <sub>2</sub> (VH0101)CH(VH0202)	4	C(H00)(CH <sub>3</sub> )(CH <sub>3</sub> )CH <sub>3</sub>	4
CH <sub>3</sub> CH(H11)(VH0303)	4	CH <sub>3</sub> CH <sub>2</sub> (VH0406)	4
CH <sub>2</sub> =CH(VH0202)	4		

**Scheme 1.** Example of a Starting Rule

**If**  
 there are no isochronous C atoms  
**AND** the number of H atoms in <sup>1</sup>H-NMR exceeds by 1 that deduced from the <sup>13</sup>C-NMR spectrum  
**AND** there is only one N and no other heteroatom in the molecular formula  
**then**  
 there are no N and NH<sub>2</sub> basic fragments (which, therefore, are bad-list entries), and NH is a good-list item

process, the dimension of the connectivity matrix is reduced and further information can be detected by interactions with other rules (see below). In addition, explicitly defined bad-list items may accelerate the structure generation process.

**Good-List Rules.** The 37 good-list rules implemented so far are summarized in Table 2. The characters within the symbols "<" and ">" specify the neighborhood of the preceding basic fragment in terms of the minimal and maximal numbers of directly bonded H atoms (e.g., H11), of vicinal H atoms (e.g., VH0406), or of other neighboring atoms (e.g., O11). This kind of information is directly used by the structure generator ASSEMBLE.<sup>17</sup> The good-list rules use NMR spectroscopic coupling and chemical shift data. Based on the results given in the preceding paper in this issue,<sup>14</sup> the 11 fragments in Table 2 were chosen so that they define as much of the chemical environment as possible. As shown earlier,<sup>18</sup> IR spectroscopy is of very limited use for creating good-list items and was, therefore, not considered here.

Whenever possible, SpecInt tries to allocate specific basic fragments to a substructure found as a good-list entry, i.e., to derive local good-list items. Such local assignments may help to detect further information (see above). Depending on the circumstances, one or several atoms of a substructure can be assigned unequivocally to basic fragments. This is the reason why up to four rules were developed for a given substructure. The examples in Scheme 2 show the various possibilities of good-list rules for an ethyl group. The first entry (a) provides the strongest evidence because it allows one to locally assign both CH<sub>3</sub> and CH<sub>2</sub>. The next two rules, b and c, give a local statement for one CH<sub>2</sub> and CH<sub>3</sub>, respectively, and rule d claims the presence of the substructure only globally without making any specific assignments.

**Bad-List Rules.** At present, SpecInt contains global bad-list rules for 45 fragments (Table 3). Additional local bad-list rules are available for each of the substructures that consist of up to two basic fragments. Local rules apply the same <sup>13</sup>C-NMR conditions as global ones, but they handle every pair of basic fragments separately. Both the local and global bad-list information is stored in the hyperstructure connectivity matrix. The possibility of globally excluding a fragment decreases with an increasing number of NMR signals. In such situations, local bad-list rules gain importance. For example, if a compound contains an ethyl group and an additional methyl with a chemical shift above 25 ppm, a local bad-list rule does not allow this methyl to be attached

**Scheme 2.** Four Good-List Rules Claiming the Presence of an Ethyl Group That Has No Neighboring Protons<sup>a</sup>

- a. If**  
 in <sup>13</sup>C-NMR  
   1 CH<sub>3</sub>  
   AND 1 CH<sub>2</sub>  
 AND in <sup>1</sup>H-NMR  
   triplet with integral x:3 and J<sub>1</sub>  
   AND quartet with integral x:2 and J<sub>2</sub>  
   AND J<sub>1</sub> = J<sub>2</sub>  
**then**  
 CH<sub>3</sub> is connected to CH<sub>2</sub>: *consequences in the connectivity matrix*  
 CH<sub>2</sub> is connected to CH<sub>3</sub> and to no other H-carrying basic fragment (except OH, NH):  
   *consequences in the connectivity matrix*  
 triplet couples only with quartet: *consequences in the HH-correlation matrix*  
 quartet couples only with triplet: *consequences in the HH-correlation matrix*  
 triplet corresponds to CH<sub>3</sub>: *consequences in the CH-correlation matrix*  
 quartet corresponds to CH<sub>2</sub>: *consequences in the CH-correlation matrix*  
 there is a CH<sub>3</sub>CH<sub>2</sub><H22><VH0303>: *good-list entry*
- 
- b. If**  
 in <sup>13</sup>C-NMR  
   more than 1 CH<sub>3</sub>  
   AND 1 CH<sub>2</sub>  
 AND in <sup>1</sup>H-NMR  
   triplet with integral x:3 and J<sub>1</sub>  
   AND quartet with integral x:2 and J<sub>2</sub>  
   AND J<sub>1</sub> = J<sub>2</sub>  
**then**  
 CH<sub>2</sub> is connected to one of the CH<sub>3</sub> and to no other H-carrying basic fragment (except OH, NH):  
   *consequences in the connectivity matrix*  
 triplet couples only with quartet: *consequences in the HH-correlation matrix*  
 quartet couples only with triplet: *consequences in the HH-correlation matrix*  
 triplet corresponds to one of the CH<sub>3</sub>: *consequences in the CH-correlation matrix*  
 quartet corresponds to CH<sub>2</sub>: *consequences in the CH-correlation matrix*  
 there is a CH<sub>3</sub>CH<sub>2</sub><H22><VH0303>: *good-list entry*
- 
- c. If**  
 in <sup>13</sup>C-NMR  
   1 CH<sub>3</sub>  
   AND more than 1 CH<sub>2</sub>  
 AND in <sup>1</sup>H-NMR  
   triplet with integral x:3 and J<sub>1</sub>  
   AND quartet with integral x:2 and J<sub>2</sub>  
   AND J<sub>1</sub> = J<sub>2</sub>  
**then**  
 CH<sub>3</sub> is only connected to one of the CH<sub>2</sub>: *consequences in the connectivity matrix*  
 triplet couples only with quartet: *consequences in the HH-correlation matrix*  
 quartet couples only with triplet: *consequences in the HH-correlation matrix*  
 triplet corresponds to CH<sub>3</sub>: *consequences in the CH-correlation matrix*  
 quartet corresponds to one of the CH<sub>2</sub>: *consequences in the CH-correlation matrix*  
 there is a CH<sub>3</sub>CH<sub>2</sub><H22><VH0303>: *good-list entry*
- 
- d. If**  
 in <sup>13</sup>C-NMR  
   more than 1 CH<sub>3</sub>  
   AND more than 1 CH<sub>2</sub>  
 AND in <sup>1</sup>H-NMR  
   triplet with integral x:3 and J<sub>1</sub>  
   AND quartet with integral x:2 and J<sub>2</sub>  
   AND J<sub>1</sub> = J<sub>2</sub>  
**then**  
 triplet couples only with quartet: *consequences in the HH-correlation matrix*  
 quartet couples only with triplet: *consequences in the HH-correlation matrix*  
 there is a CH<sub>3</sub>CH<sub>2</sub><H22><VH0303>: *good-list entry*

<sup>a</sup> Rule a allows local assignment of both CH<sub>3</sub> and CH<sub>2</sub>, rules b and c locally assign only one of these groups, and rule d only globally claims their presence.

to a methylene, whereas an ethyl group, evidently, cannot be forbidden globally.

Scheme 3 shows a bad-list rule for excluding a CH<sub>3</sub>CH group. As in most bad-list rules, the <sup>1</sup>H-NMR coupling information is used because it provides the strongest evidence for fragments having several free valences. In this example, only the lower limit of the number of lines in the multiplet is required since the number of vicinal H atoms of the CH group is not defined. In addition, the methine proton can exhibit a very large chemical shift range because it could have electronegative substituents and the C atom might be sp<sup>2</sup>-hybridized. The corresponding <sup>13</sup>C-NMR shift range is

so broad that no constraint was used at all. The four IR ranges given correspond to the stretching, asymmetric, and symmetric deformation and the rocking vibration of the methyl group.

As shown in Schemes 2 and 3, the different conditions are connected with a logical AND in good-list and with OR in bad-list rules. Hence, in the case of good-list rules, all conditions must be fulfilled simultaneously. Their reliability is thus enhanced, which would not be possible if the various spectroscopic methods, as in most other systems, were treated successively. As a consequence of the simultaneous treatment, it is impossible to draw a conclusion about a good-

**Table 3.** Fragments for Which Global Bad-List Rules Are Available in SpecInt<sup>a</sup>

C≡C	C(=O)OH	NH=N	CH <sub>3</sub> N
HC≡C	C(=O)OC	NH <sub>2</sub>	CH <sub>3</sub> NH
C=C	CC(=O)C	C≡N	CH <sub>3</sub> NH <sub>2</sub>
CH=C	C-O	C=N	CHNH <sub>2</sub>
CH=CH	CH <sub>2</sub> O	C=NH	CH <sub>2</sub> NH <sub>2</sub>
CH <sub>2</sub> =C	CH <sub>3</sub> O	CH=N	C(=O)NH <sub>2</sub>
CH <sub>2</sub> =CH	CH <sub>3</sub> CH <sub>2</sub> O	CH=NH	CH <sub>2</sub> F
CH <sub>3</sub> CH	COC	CH <sub>2</sub> =N	NOH
CH <sub>3</sub> CH <sub>2</sub>	C-OH	C-N	N(=O)O
C=O	CHOH	CNH	S(=O)=O
CH=O	CH <sub>2</sub> OH	CHNH	
C(=O)O		CH <sub>2</sub> NH	

<sup>a</sup> For those consisting of less than three basic fragments, local bad-list rules have also been implemented.

**Scheme 3.** Example of a Bad-List Rule

**If**  
 in <sup>13</sup>C-NMR  
 no signal for CH<sub>3</sub> between 2.4 and 32.1 ppm  
**OR** in <sup>1</sup>H-NMR  
 no doublet or triplet between 0.4 and 2.5 ppm  
**OR** no coupling pattern having at least 4 lines between 0.5 and 7.6 ppm  
**OR** in IR  
 no signal in at least one of the following ranges: 3095-2840, 1495-1410, 1426-1334, 1195-890 cm<sup>-1</sup>  
**then**  
 put CH<sub>3</sub>CH on the bad-list

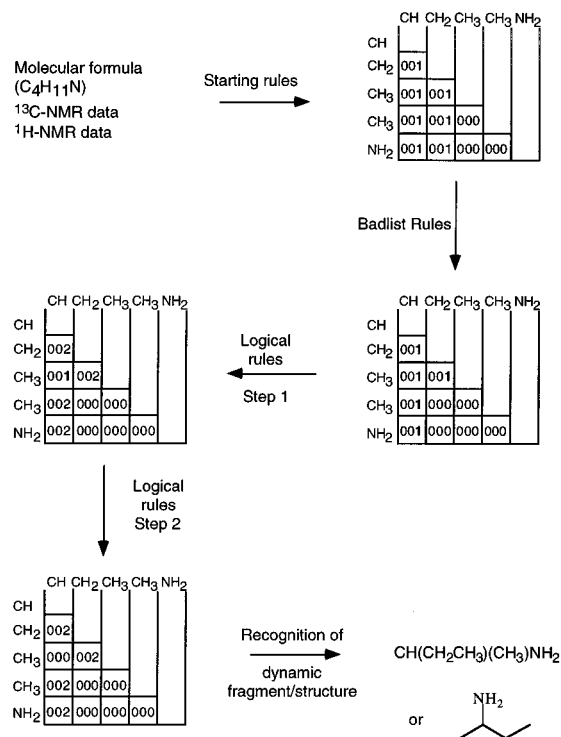
list item if one of the different spectra is not available. On the other hand, the connection of the various conditions in bad-list rules with logical OR's means that, even if only one of them is fulfilled, the corresponding rule is activated. Therefore, a simultaneous treatment of bad-list conditions is pointless. Moreover, if one spectral method is not at hand, the conditions based on the remaining ones are still valid so that only the efficiency of the rule is reduced.

**Logical Rules.** There are 19 rules that operate on the three matrices of Figure 1 in order to find implicit information. Since every matrix update may reveal new implicit information, these rules are applied repeatedly until they produce no more changes. As a simple example of a logical rule, we assume that three basic fragments are available: fragment A with four free valences and exactly two possible neighbors as well as fragments B and C in which a double and a triple bond, respectively, are forbidden. A logical rule then derives the existence of the fragment C-A≡B. Such a fragment, whose presence or absence is not explicitly stated in the good-list or bad-list rules, is called dynamic as opposed to static. One of the strengths of SpecInt is its capability finding such dynamic fragments previously unknown to the system. Part of the logical rules check these newly built fragments against the previously created static ones and detect possible overlaps. Two fragments having the same constitution, but different specifications of their neighborhoods (i.e., different atom tags) may or may not describe the same substructure. For example, CH<sub>2</sub>CH as a global good-list fragment and CH<sub>2</sub>-(H22)CH(H11) as a local one are considered to possibly overlap.

To achieve the highest reliability possible, a number of logical rules perform consistency checks. On the one hand, prior to every change in a matrix, such tests are carried out in that the system, before making a contradictory entry, stops the process and reports which rules have caused what kind of error. On the other hand, a further set of logical rules perform final tests after all rules have been run. Among

others, they search for the following contradictions: a bad-list fragment figuring as part of a good-list substructure; good-list or bad-list rule contradicting the result of a logical rule (for example, a good-list rule erroneously predicts CH<sub>3</sub>-CH<sub>2</sub> as a global fragment of the unknown compound (as mentioned above, no specific bonds can be assigned in the connectivity matrix for global goodlist entries); since logical rules may forbid certain bonds, it may happen that CH<sub>3</sub>CH<sub>2</sub> at the same time also occurs as a dynamic bad-list fragment); a basic fragment has more bonds than free valences, or these are not satisfied by the remaining bonds allowed.

**Practical Example.** A simple example is presented in Figure 2 to show the synergisms of the different kinds of

**Figure 2.** Example showing the cooperation of the various rules. In this simple case, the correct structure was derived.

rules. From the molecular formula (C<sub>4</sub>H<sub>11</sub>N), the <sup>13</sup>C-NMR data (CH<sub>3</sub> at 10.7 and 23.6 ppm, CH<sub>2</sub> at 33.0 ppm, and CH at 48.5 ppm), and the sum of the normalized integrals (11) of the <sup>1</sup>H-NMR spectrum, SpecInt first set up the connectivity matrix. The primary rules found that double and triple bonds were forbidden, that the nitrogen had to be present as NH<sub>2</sub>, no NH and N being possible as basic fragments, and that a bond between the terminating fragments CH<sub>3</sub> and NH<sub>2</sub> was excluded. The bad-list rule for CH<sub>2</sub>NH<sub>2</sub> did not find a possibly matching coupling pattern in the <sup>1</sup>H-NMR spectrum so that this fragment was globally forbidden. Since the unknown compound has an ethyl group, this substructure was, of course, not globally excluded by the bad-list rule for CH<sub>3</sub>CH<sub>2</sub>. However, after checking the basic fragments locally, a bond between the CH<sub>2</sub> and the CH<sub>3</sub> at 23.6 ppm was forbidden owing to this chemical shift value. The logical rules then established the only possible bonds left, i.e., those between NH<sub>2</sub> and CH, between the CH<sub>3</sub> at 23.6 ppm and CH, and the two bonds of the CH<sub>2</sub> basic fragment. Further application of the logical rules established the last possible bond between the CH<sub>3</sub> at 10.7 ppm and CH<sub>2</sub>. Thus, a complete structure was produced and recognized as a

**Table 4.** Reduction in the Total Number of Possible Structures Using Structural Information Provided by SpecInt


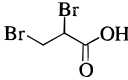
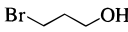

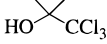
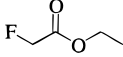



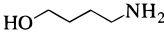


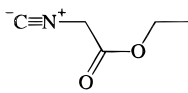

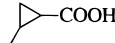




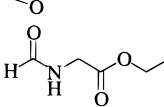




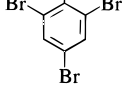

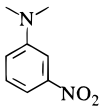
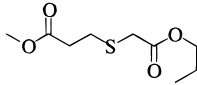
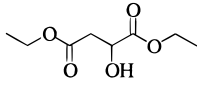
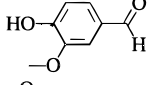
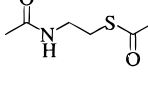
molecular formula	no. of possible structures	no. of remaining structures (reduction factor, %) generated with ASSEMBLE		correct structure
		incl <sup>a</sup>	excl <sup>a</sup>	
C <sub>3</sub> H <sub>3</sub> Cl <sub>3</sub> O	54	12 (77.8)	12 (77.8)	
C <sub>3</sub> H <sub>4</sub> Br <sub>2</sub> O <sub>2</sub>	201	13 (93.5)	13 (93.5)	
C <sub>3</sub> H <sub>7</sub> BrO	10	1 (90.0)	1 (90.0)	
C <sub>4</sub> H <sub>6</sub> O <sub>2</sub>	263	9 (96.7)	7 (97.3)	
C <sub>4</sub> H <sub>7</sub> Cl <sub>3</sub> O	108	2 (98.1)	2 (98.1)	
C <sub>4</sub> H <sub>7</sub> FO <sub>2</sub>	487	9 (98.2)	9 (98.2)	
C <sub>4</sub> H <sub>8</sub> N <sub>2</sub> O	6 754	103 (98.5)	101 (98.5)	
C <sub>4</sub> H <sub>9</sub> BrO <sub>2</sub>	115	1 (99.1)	1 (99.1)	
C <sub>4</sub> H <sub>11</sub> N	8	1 (87.5)	1 (87.5)	
C <sub>4</sub> H <sub>11</sub> NO	211	1 (99.5)	1 (99.5)	
C <sub>5</sub> H <sub>4</sub> O <sub>2</sub>	1 821	22 (98.8)	7 (99.6)	
C <sub>5</sub> H <sub>4</sub> O <sub>3</sub>	7 744	373 (95.2)	118 (98.5)	
C <sub>5</sub> H <sub>7</sub> NO <sub>2</sub>	44 336	197 (99.6)	117 (99.7)	
C <sub>5</sub> H <sub>8</sub> O <sub>2</sub>	1 168	10 (99.1)	10 (99.1)	
C <sub>5</sub> H <sub>8</sub> O <sub>2</sub>	1 168	3 (99.7)	2 (99.8)	
C <sub>5</sub> H <sub>8</sub> O <sub>2</sub>	1 168	5 (99.6)	4 (99.7)	
C <sub>5</sub> H <sub>8</sub> O <sub>2</sub>	1 168	14 (98.8)	12 (99.0)	
C <sub>5</sub> H <sub>9</sub> NO	3 390	28 (99.2)	27 (99.2)	
C <sub>5</sub> H <sub>9</sub> NO	3 390	5 (99.9)	5 (99.9)	
C <sub>5</sub> H <sub>9</sub> NO <sub>3</sub>	109 126	1 367 (98.7)	1 285 (98.8)	
C <sub>5</sub> H <sub>10</sub> O	74	3 (95.9)	3 (95.9)	
C <sub>5</sub> H <sub>12</sub> O	14	1 (92.9)	1 (92.9)	
C <sub>5</sub> H <sub>10</sub> O <sub>2</sub>	400	14 (96.5)	14 (96.5)	
C <sub>5</sub> H <sub>13</sub> N	17	1 (94.1)	1 (94.1)	
C <sub>6</sub> H <sub>3</sub> Br <sub>3</sub> O	19 969	306 (95.9)	77 (99.6)	
C <sub>6</sub> H <sub>5</sub> Br	685	26 (96.2)	7 (99.0)	

Table 4 (Continued)

molecular formula	no. of possible structures	no. of remaining structures (reduction factor, %) generated with ASSEMBLE		correct structure
		incl <sup>a</sup>	excl <sup>a</sup>	
C <sub>6</sub> H <sub>5</sub> FS	8 372	73 (99.1)	17 (99.8)	
C <sub>6</sub> H <sub>6</sub> OS	28 521	857 (97.0)	377 (98.7)	
C <sub>6</sub> H <sub>8</sub> O <sub>3</sub>	54 343	1 166 (97.9)	1 066 (98.0)	
C <sub>6</sub> H <sub>9</sub> BrO <sub>3</sub>	135 088	1 795 (98.7)	1 740 (98.7)	
C <sub>6</sub> H <sub>9</sub> NO <sub>2</sub>	272 736	521 (99.8)	298 (99.9)	
C <sub>6</sub> H <sub>9</sub> NO <sub>2</sub>	272 736	1 553 (99.4)	1 448 (99.5)	
C <sub>6</sub> H <sub>10</sub> O <sub>2</sub>	4 869	60 (98.8)	51 (99.0)	
C <sub>6</sub> H <sub>10</sub> O <sub>2</sub>	4 869	37 (99.2)	32 (99.3)	
C <sub>6</sub> H <sub>10</sub> O <sub>3</sub>	23 838	145 (99.4)	135 (99.4)	
C <sub>6</sub> H <sub>12</sub> N <sub>2</sub> O <sub>3</sub>	7 822 063	97 (99.999)	50 (99.999)	
C <sub>6</sub> H <sub>12</sub> O	211	6 (97.2)	6 (97.2)	
C <sub>6</sub> H <sub>12</sub> O	211	6 (97.2)	6 (97.2)	
C <sub>6</sub> H <sub>12</sub> O <sub>2</sub>	1 313	30 (97.7)	30 (97.7)	
C <sub>7</sub> H <sub>10</sub> O <sub>3</sub>	308 660	1 267 (99.9)	995 (99.6)	
C <sub>7</sub> H <sub>14</sub> O <sub>3</sub>	22 151	440 (98.0)	440 (98.0)	
C <sub>7</sub> H <sub>14</sub> O <sub>3</sub>	22 151	38 (98.9)	38 (98.9)	
C <sub>7</sub> H <sub>15</sub> NO <sub>2</sub>	86 195	214 (99.8)	214 (99.8)	
C <sub>8</sub> H <sub>14</sub> O <sub>4</sub>	2 224 538	41 (99.9)	39 (99.9)	
C <sub>8</sub> H <sub>16</sub> O	1 684	13 (99.2)	13 (99.2)	
C <sub>8</sub> H <sub>19</sub> ClSi	1 608	6 (99.6)	6 (99.6)	
C <sub>9</sub> H <sub>12</sub> O	338 761	338 (99.9)	133 (99.9)	
C <sub>9</sub> H <sub>15</sub> NO	3 430 261	961 (99.97)	602 (99.98)	
C <sub>10</sub> H <sub>14</sub> O	1 548 361	212 (99.9)	85 (99.9)	

Table 4 (Continued)

molecular formula	no. of possible structures	no. of remaining structures (reduction factor, %) generated with ASSEMBLE		correct structure
		incl <sup>a</sup>	excl <sup>a</sup>	
C <sub>8</sub> H <sub>10</sub> N <sub>2</sub> O <sub>2</sub> <sup>b</sup>	808 891 281	6 960 563 (99.1)		
C <sub>9</sub> H <sub>16</sub> O <sub>4</sub> S <sup>b</sup>	217 879 071	418 460 (99.8)		
C <sub>8</sub> H <sub>14</sub> O <sub>5</sub> <sup>b</sup>	9 596 344	104 199 (98.9)		
C <sub>8</sub> H <sub>8</sub> O <sub>3</sub> <sup>b</sup>	6 333 319	153 810 (97.6)		
C <sub>6</sub> H <sub>11</sub> NO <sub>2</sub> S <sup>b</sup>	1 641 587	5 842 (99.6)		

<sup>a</sup> Including or excluding. <sup>b</sup> Generated with MOLGEN version 3.1.

dynamic one. Other rules, e.g., the (global) good-list rules predicting CH<sub>3</sub>CH and CH<sub>3</sub>CH<sub>2</sub> to be present, although they did not contribute any new information, confirmed the structure of the 2-aminobutane found.

**Tests.** The reliability of the good-list and bad-list rules was tested against 99 059 <sup>13</sup>C-NMR,<sup>19</sup> 8000 <sup>1</sup>H-NMR,<sup>20</sup> and 34 229 IR spectra,<sup>19</sup> and all rules were found to be valid without exception. Naturally, this 100% reliability regarding the databases used does not guarantee against violating a rule and missing a valid structure. However, none of this happened in any of the 100 test cases, 54 of which are presented in Table 4. Except for the last 5, for which version 3.1 of MOLGEN<sup>21</sup> was applied, structures were generated with ASSEMBLE because it makes almost full use of all of the information available.<sup>14</sup> The mean reduction factor of the solution space was 97.5% (97.8%) and the corresponding median 98.8% (99.0%), the values in parentheses being obtained when chemically improbable solutions were excluded.<sup>17</sup> Owing to the fact that the solution space is quite small in many cases, the results are biased toward lower percentages. In six examples, the automatic spectra interpretation routines led to a single solution. In none of the 100 cases investigated was the correct structure missed. It must be kept in mind, though, that the results refer to small- and medium-sized molecules. For larger ones, it was not possible to investigate the complete solution space. It is expected that the performance of SpecInt decreases with increasing molecule size because local good-list and global bad-list predictions will be more difficult. However, an extension of the rules is planned. At the present stage, mass spectral data have not yet been included. They would be useful, especially for good-list entries, for which large fragments with few free valences are efficient. Possibly, a combination of existing mass spectral rules (MSClass<sup>13</sup>) with other spectroscopic information could provide a set of highly reliable good-list predictions. Of course, less conservative rules could also improve the performance but would increase the risk of missing the valid structure.

#### 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

- 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.
- 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.
- Dubois, J. E.; Carabedian, M.; Dagane, I. Computer-aided elucidation of structures by carbon-13 NMR. *Anal. Chim. Acta* **1984**, *158*, 217–233.
- 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.
- Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra. 6. Substructure constraints from analysis of <sup>13</sup>C-NMR spectra. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 979–1000.
- 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.
- 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.
- 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.
- Gray, N. A. B. Structural interpretation of spectra. *Anal. Chem.* **1975**, *47*, 2426–2431.
- Luinge, H.-J.; Maas, J. H. v. d. Artificial intelligence for the interpretation of combined spectral data. *Anal. Chim. Acta* **1989**, *223*, 135–147.
- 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.
- Woodruff, H. B.; Smith, G. M. Computer program for the analysis of infrared spectra. *Anal. Chem.* **1980**, *52*, 2321–2327.

- (13) Varmuza, K.; Werther, W. Mass spectral qualifiers for supporting systematic structure elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323–333.
- (14) Schriber, H.; Pretsch, E. General characteristics of good-list and bad-list entries for structure generators from spectra. *J. Chem. Inf. Comput. Sci.*, preceding paper in this issue.
- (15) 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.
- (16) Bohanec, S. Structure generation by the combination of structure reduction and structure assembly. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 494–503.
- (17) Shelley, C. A.; Munk, M. E. CASE, a computer model of the structure elucidation process. *Anal. Chim. Acta* **1981**, *133*, 507–516.
- (18) Affolter, C.; Baumann, K.; Clerc, J. T.; Schriber, H.; Pretsch, E. Automatic interpretation of infrared spectra. *Mikrochim. Acta [Suppl.]* **1997**, *14*, 143–147.
- (19) *SpecInfo*; Chemical Concepts GmbH: Weinheim, Germany, 1993.
- (20) Sasaki, S. *Handbook of Proton-NMR Spectra and Data*; Academic Press: London, 1987.
- (21) Wieland, T.; Kerber, A.; Laue, R. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413–419.

CI970014X