Validation of Structural Proposals by Substructure Analysis and ^{13}C NMR Chemical Shift Prediction

Jens Meiler,[†] Erdogan Sanli,[†] Jochen Junker,[†] Reinhard Meusinger,[‡] Thomas Lindel,[§] Martin Will,[||] Walter Maier,[||] and Matthias Köck*,[⊥]

Alfred-Wegener-Institut für Polar- und Meeresforschung, Am Handelshafen 12, D-27570 Bremerhaven, Germany, Institut für Organische Chemie, J. W. Goethe-Universität Frankfurt, Marie-Curie-Strasse 11, D-60439 Frankfurt, Germany, Institut für Organische Chemie, Universität Mainz, D-55099 Mainz, Germany, Department Chemie, Universität München, Butenandtstrasse 5-13, D-81377 München, Germany, and BASF AG, D-67056 Ludwigshafen, Germany

Received March 20, 2001

The 2D NMR-guided computer program COCON can be extremely valuable for the constitutional analysis of unknown compounds, if its results are evaluated by neural network-assisted ^{13}C NMR chemical shift and substructure analyses. As instructive examples, data sets of four differently complex marine natural products were thoroughly investigated. As a significant step towards a true automated structure elucidation, it is shown that the primary COCON output can be safely diminished to less than 1% of its original size without losing the correct structural proposal.

INTRODUCTION

NMR-based structure generators are of special importance for the constitutional analysis of underdetermined proton-poor compounds. Frequently, a very large number of constitutions is in accordance with the NMR correlation data for such systems. Therefore, computer-assisted methods are required to validate these results. Recently,[1] we have demonstrated that the calculation of the ^{13}C NMR chemical shifts (δ(^{13}C)) with the HOSE code based program SpecEdit[2] is important for the evaluation of structural proposals. The difference between the experimental and the theoretical values ($\Delta[\delta(^{13}\mathrm{C})]$) is very useful for the ranking of the structural proposals. For large data sets (>10 000 structural proposals) these calculations are rather time-consuming because of an approximate calculation time of 1 s per structure (calculation times >3 h).

In this contribution two approaches to solve this problem are presented (see Figure 1):

(a) An acceleration in the calculation of ^{13}C NMR chemical shifts (δ(^{13}C)). A neural network approach is used to ensure a fast and accurate chemical shift prediction of the constitutions generated by COCON (*Co*nstitutions from *con*nectivities).[3] Neural networks have become an effective method in chemistry as a flexible tool for data handling and analysis. Several examples of neural networks were already published for the analysis[4] and the prediction of NMR spectra.[5]

(b) Substructure analysis. A new implementation of a substructure analysis based on the comparison of atomic environments will be introduced. A substructure analysis allows one to investigate the diversity of a set of structural proposals. Algorithms for determining common substructures in a set of structural proposals are still the subject of development.[6]

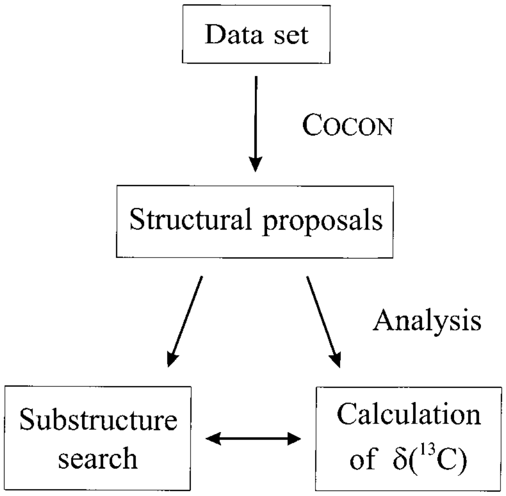

Figure 1. General scheme for a systematic analysis of natural products based on NMR spectroscopic data as presented in this contribution.

To demonstrate the efficiency of these approaches, several sets of structural proposals generated by COCON[7] were investigated. COCON uses connectivity information from two-dimensional NMR spectroscopy to generate all possible structures of a molecule which agree with this information for a given molecular formula. It is demonstrated that both comparison of calculated δ(^{13}C) with experimental data and substructure analysis are efficient tools to perform the validation of these structural proposals.

METHODS

The data sets of four marine natural products of differing complexity were used as input for COCON calculations. This

* Corresponding author phone: +49-471-4831-1497; fax: +49-471-4831-1425; e-mail: mkoeck@awi-bremerhaven.de.

[†] Institut für Organische Chemie, Universität Frankfurt.

[‡] Institut für Organische Chemie, Universität Mainz.

[§] Department Chemie, Universität München.

[||] BASF AG, Ludwigshafen.

[⊥] Alfred-Wegener-Institut für Polar-und Meeresforschung, Bremerhaven.

10.1021/ci010294n CCC: $22.00 © 2002 American Chemical Society
Published on Web 03/08/2002

Table 1. Results of the COCON and $\Delta[\delta(^{13}C)]$ Calculations for Compounds **1**−**4**

	COCON				$\Delta[\delta(^{13}C)]$		
	data[a]	correlation[b]	no.[c]	calc time[d]	range[e]	best[f]	calc time[g]
oroidin (**1**)	exptl	111000	33	0.3 s	7.8−20.5	9.1 (2)	<1 s
manzacidin (**2**)	E + T	111000	190	3.9 s	5.2−27.1	5.2 (1)	<1 s
5-deoxyenterocin (**3**)	exptl	110000	82	4.9 s	3.7−21.7	3.7 (1)	<1 s
ascididemin (**4**)	theor	110000	28 672	1 min 43 s	4.9−29.5	6.7 (25)	7 min 56 s[h]

[a] Origin of the correlation data set: exptl (E) stands for experimental, theor (T) for theoretical (explanation, see text). [b] Correlation data used for the COCON calculations. The six columns stand for $^1H,^1H$ COSY, $^1H,^{13}C$ HMBC, 1,1-ADEQUATE, $^1H,^{15}N$ HMBC, fixed and forbidden (1 indicates that the data is used and 0 that it is not used). [c] Number of structural proposals generated by COCON under consideration of the correlation data given in the correlation column. [d] Calculation times were obtained with a Silicon Graphics R10000 processor. The COCON source code was 64-bit compiled. [e] Range of the $\delta(^{13}C)$ deviations [ppm] (calculated − experimental) for all structural proposals. [f] $\delta(^{13}C)$ deviation [ppm] for the best structural proposal. The ranking of the correct strcuture is given in parentheses. [g] Calculation times were obtained on a PC Pentium II, 450 MHz. [h] This value is the totel calculation time (including reading of the files). The pure chemical shift calculation is 103 s. The neural network is able to calculate 5000 ^{13}C chemical shifts per second. The calculation time for SpecEdit was 300 min (about 30 ^{13}C chemical shifts per second). The ranking is comparable, although the absolute values are better for SpecEdit. The correct constitution has a $\Delta[\delta(^{13}C)]$ of 1.2.

resulted in the generation of 33 (compound **1**) to 28 672 (compound **4**) structural proposals. In all cases, the correct structure is ranked within the first 0.2% in a hit list of all structural proposals (see Table 1) by calculating the $\delta(^{13}C)$ deviations between the experimental and the theoretical values ($\Delta[\delta(^{13}C)]$).

The first two examples are bromopyrrole alkaloids from marine sponges. Both natural products oroidin (**1**)[8] and manzacidin A (**2**)[9] were used as model compounds for COCON calculations before.[7a,b] The correlation data for **1** is described in ref 7b; for **2** the original data from 1991 was used as input for the COCON analysis. 5-Deoxyenterocin (**3**) was isolated from a tunicate of the genus *Didemnum* in 1996.[10] The published correlation data served as input for the COCON calculation. Ascididemin (**4**), a pyridoacridine alkaloid, was first isolated in 1988 from the tunicate *Didemnum* species.[11] It represents an example of a proton-poor compound for which only $^1H,^1H$ COSY and $^1H,^{13}C$ HMBC correlations are available. A theoretical data set was used for this example. Theoretical data set means that all $^3J_{HH}$ and $^2J_{CH}/^3J_{CH}$ correlations of the given constitution of ascididemin (**4**) were extracted.

CALCULATION OF ^{13}C CHEMICAL SHIFT

The inclusion of ^{13}C NMR chemical shifts as orthogonal (not correlated) information to the connectivity constraints used by the structure generator COCON optimizes the efficiency of a subsequent analysis of the resulting structural proposals. This leads to large deviations between the experimental and the predicted chemical shifts for many carbons in the generated structures. The resulting wide distribution of $\Delta[\delta(^{13}C)]$ deviations provides an effective filter. A fast and accurate method for determining ^{13}C chemical shifts of organic substances is available using artificial neural networks.[3] So far the ^{13}C chemical shift prediction was carried out using large computer-stored databases or incremental methods. Both methods rely on a spherical encoding (introduced by Bremser[12]) of the environment of a carbon atom. While databases such as Specinfo,[13] SpecEdit,[2] and CSearch[14] provide an accurate shift prediction, they have rather long calculation times (although there are approaches to accelerate these searches[15]) and their availibility and flexibility suffer due to the dependence on the direct access to the large amount of data. Incremental methods[16] are usually very fast but lead to large deviations for complex structures because interactions between substituents are not considered here.[5g] Artificial neural networks allow one to calculate ^{13}C chemical shifts nearly as fast as incremental methods, which is still about 100 times faster in comparison to a database search with no loss in accuracy.[3]

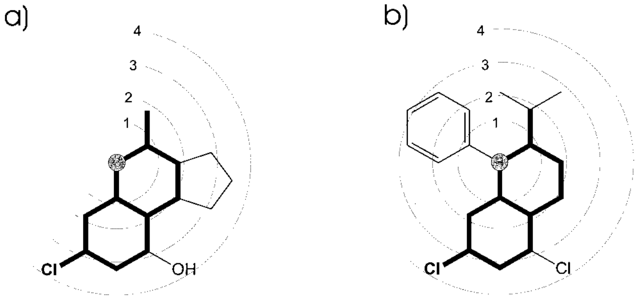

Figure 2. Largest common substructure of two model compounds. The highlighted carbon atoms are superimposed first. Subsequently, the atoms in the increasing numbered spheres are compared and added to the substructure until no further superimposition is possible. The largest substructure is bold marked.

From the spherical environment of a carbon atom (see Figure 2) a numerical code is derived containing the number of atoms, their atom type, and the hybridization state. The environment of a carbon atom can be subdivided into spheres. This is carried out by counting the minimal number of bonds between the carbon atom of interest and every other atom, respectively. The first five spheres and an additional sum sphere (which considers all atoms from the sixth to higher spheres) are taken into consideration. All atoms are subdivided in 28 atom types according to their atomic number, hybridization state, and the number of attached protons. In each sphere the frequency of atoms for every atom type (28), the number of protons, and the number of ring closures are determined (+2). This is carried out first for all atoms and in the next step only for atoms belonging to a conjugated π-electronic system including the carbon atom of interest. Therefore, the environmental code of a single carbon atom consists of (28 + 2) parameters for each of the six spheres and for two separate groups of σ- and π-bonded atoms, respectively, which leads to an overall 360 numbers (30 times 6 times 2). Nine out of the 28 defined atom types are carbon atoms which are defined as follows: (a) four aliphatic (⟩C⟨ , ⟩CH−, $-CH_2-$, $-CH_3$), (b) two olefinic (=C⟨ , =CH−, or $=CH_2$), (c) one triple bonded (≡C− or ≡CH or =C=) and (d) two aromatic ( ) ⟩C−,) ⟩CH). For each of them an individual neural network was established which uses a vector with 360 numbers as input and predicts the chemical shift. After training these neural networks with 40 000 compounds from the Specinfo database, the average deviation of the ^{13}C NMR chemical shift calculation was determined to be 1.8 ppm for an independent data set of 5000 molecules (depending on the atom type and the hybridization state of the carbon atom).[3] These 40 000 compounds represent over 500 000 carbon atoms with a contribution of 4%, 9%, 19%, 15%, 11%, 6%, 1%, 14%, and 21% with respect to the nine carbon atom types (see above).

^{13}C chemical shifts are already considered by Cocon in the structure generation process but only on a very basic level. The ^{13}C chemical shift rules of COCON are as follows: (a) C=S and C=O bonds are forbidden if $\delta_C < 150$ ppm, (b) aliphatic C—O bonds are forbidden if $\delta_C < 45$ ppm, (c) olefinic C−O bonds are forbidden if $\delta_C < 130$ ppm, (d) olefinic C−N bonds are forbidden if $\delta_C < 105$ ppm, and (e) methyl−C bonds are forbidden if $\delta_{CH_3} > 35$ ppm.

SUBSTRUCTURE ANALYSIS

Computer programs such as COCON often generate similar structures with equivalent basic structural elements (e.g., closed ring systems) but with a different arrangement of substituents. To separate the information, a substructure analysis is of special interest. This allows one to investigate a small number of basic common substructures and the different substitution patterns. For a chemist it would be very time-consuming to perform this analysis by hand if the data set is large. However, it would be an important information to find, e.g., 10 common substructures out of 500 generated constitutions.

Furthermore, this analysis can be easily combined with a ^{13}C chemical shift calculation in two ways:

(a) Only the generated structures with the smallest ^{13}C chemical shift deviations ($\Delta[\delta(^{13}C)]$) to the experimental data are used for the substructure analysis. This might become necessary if the number of generated constitutions is too large to perform a full substructure analysis or the resulting set of substructures would become too complex for further investigations.

(b) It is possible to calculate an average chemical shift value for every carbon atom in a substructure. This is carried out by averaging the chemical shift values of the corresponding carbon atoms in molecules which contain this particular substructure. As will be shown later, this averaging leads to smaller deviations of the chemical shift to the experimental values, if the substructure is a part of the correct structure.

The set of substructures is calculated by combining all structural proposals pairwise (see Figure 3). For every pair of molecules the largest common substructure is computed. A substructure of two molecules is defined such that all superimposed atoms within the substructure are (a) of the same element type (C, N, O, ...) and (b) are connected by exactly the same bond types (single, double, triple, or aromatic). If the investigated ensemble contains n molecules, $n(n - 1)/2$ substructures have to be generated. If necessary the number of proposals considered for the substructure analysis can be limited to the molecules with the lowest $\Delta[\delta(^{13}C)]$ to the experimental values.

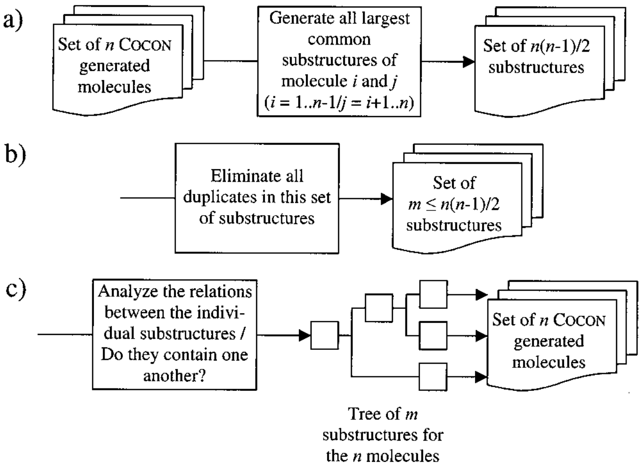
a)
Set of n COCON generated molecules
Generate all largest common substructures of molecule i and j (i = 1..n-1/j = i+1..n)
Set of n(n-1)/2 substructures
b)
Eliminate all duplicates in this set of substructures
Set of m ≤ n(n-1)/2 substructures
c)
Analyze the relations between the individual substructures / Do they contain one another?
Set of n COCON generated molecules
Tree of m substructures for the n molecules

Figure 3. Simplified flow chart diagram for the generation of the substructure tree. (a) First, for every pair of two molecules out of the set of COCON structures, the largest common substructure is computed. (b) In a second step all duplicates are eliminated. (c) Now the relations between the substructures are analyzed and the substructure tree is built. Besides the relations between several substructures, this tree also refers to members of the original set of molecules that contain the particular substructure.

Every substructure is taken into consideration only once. A newly generated substructure is tested if it is part of any structural proposal. Every generated substructure refers to all molecules and also to all other substructures that contain this particular fragment. The key function of this analysis is a procedure that generates the largest common substructure from two given structures. The largest common substructure can be found by an algorithm that associates atoms of the first structure with atoms of the second structure. Two atoms can be potentially associated if they have the same atomic number and are connected to all other atoms of the new substructure by identical bond types. Hydrogen atoms are not taken into consideration explicitly. Due to this definition, more than one association can be usually found for two molecules. The association with the maximal number of atoms is the largest common substructure. Figure 2 gives two molecules with a bold-marked largest common substructure as example.

The problem to find the largest possible association is a tree search type analysis in a mathematical sense. Nodes of two trees have to be assigned to each other. Figure 2 illustrates this problem. Similar to the earlier discussed spherical definition of an atomic environment, a recursive function is used for this purpose which starts from one atom and compares its environment sphere by sphere with the environment of another atom. Two atoms of the same atomic number are selected from both molecules and superimposed to become the first part of the new substructure. Its neighbors are assigned now sphere by sphere. If the element type (C, N, O, ...) and the bond type (single, double, triple or aromatic), are equivalent, the atom is added to the substructure. The substructure increases until no further superimposition is possible.

However, some special problems have to be considered performing such an assignment of two structures. The selection of the two starting atoms influences the result of the procedure and has therefore to be changed incrementally over all possible atom−atom combinations in an outer loop.

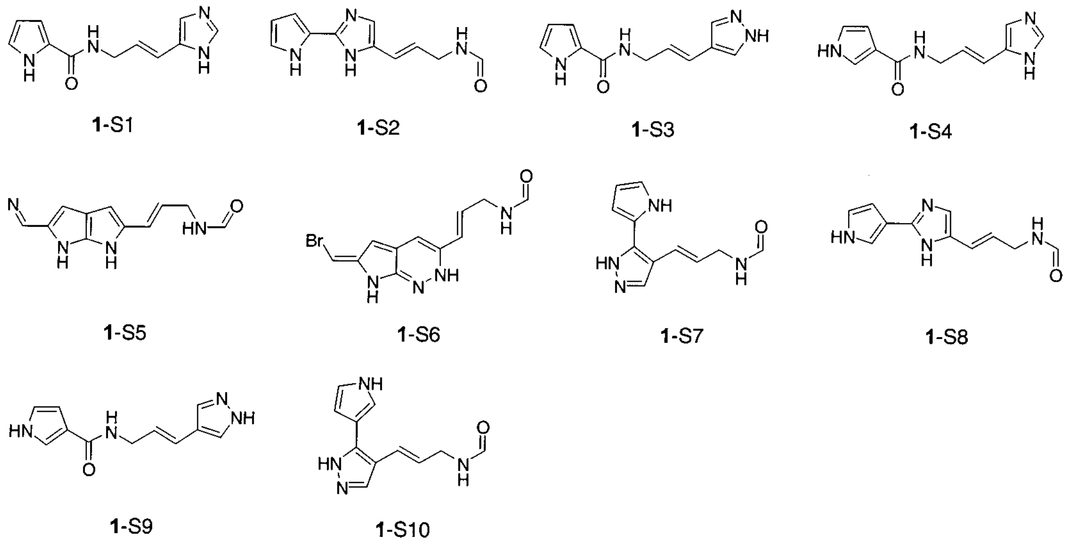

Figure 4. Results for the substructure analysis (output as given by the program) carried out for the 33 structural proposals of oroidin (**1**). A minimum of 15 atoms per substructure as well as a reorganization in a tree was applied. The variable atoms, the two bromines and the amino group, are not explicitly given in the substructures, nor is the variable proton which is always bound to the pyrrole ring (as an example, see Figure 6).

Furthermore, more than one possibility in the recursive sphere-by-sphere assignment can occur and all possibilities have to be tested in these cases. This procedure is tree based and has therefore to consider the mathematically special case of ring closures within this tree.

During the development and the testing of this procedure, it turned out that additional options are necessary which allow the generation of “intelligent” sets of substructures and limit their number. Therefore, several options (O1−O6) are introduced:

(O1) definition of a minimum number of atoms in a substructure

(O2) definition of a minimum number of rings in a substructure, to prefer substructures that include large closed ring systems (acyclic substructures may not be helpful for polycyclic molecules)

(O3) definition of a maximum number of “non ring atoms” in a substructure, to prefer substructures with large closed ring systems without substituents

(O4) definition of a minimum number of molecules per substructure, to find substructures that are common in many of the generated structures

(O5) analysis of only a part of all structures (for example, the first 1% of the structural proposals with the lowest deviation from the experimental ^{13}C NMR spectrum) to reduce the number of the generated substructures

(O6) generation of reduced sets of substructures by selecting a small set of substructures out of all generated substructures

This selection (O6) is preformed in order to find the smallest “complete” set of substructures which covers every generated molecule with exactly one substructure.

Several options for the visualization of the substructure analysis are introduced and used for the described problems (V1−V4):

(V1) The results are sorted by their averaged deviation from the ^{13}C NMR spectrum, to rank the substructures according to their probability of occurrence in the correct structure.

(V2) The results are sorted by the number of atoms in the substructures, to rank the substructures according to their size.

(V3) The results are sorted by the number of molecules that contain a particular substructure, to rank the substructures according to their frequency of occurrence.

(V4) The substructures are reorganized as a tree. This reorganization is performed by validating the relations between the substructures and by testing if a substructure is part of another substructure. The result is a plot which starts with small substructures in a first generation. All substructures containing this small substructure are given in a second generation and so on until the last generation of substructures is reached and the generated structures that contain these substructures are given. This tree or a part of it allows analysis of the relations between the substructures (see Figures 7, 8 and 10).

The ^{13}C NMR chemical shift calculation as well as the substructure analysis are combined in the program “Analyze”.[17]

RESULTS AND DISCUSSION

The calculation times for COCON and the ^{13}C chemical shift prediction of the four compounds are given in Table 1. Oroidin (**1**) was already discussed in the literature.[7b] It is used here to demonstrate both approaches on a small set of structural proposals. The results of the substructure analysis for **1** can therefore be validated by hand, allowing the approach to be tested and optimized. COCON generates 33 structural proposals for the experimental data set of oroidin (**1**) including 6 $^{1}H,^{1}H$ COSY, 23 $^{1}H,^{13}C$ HMBC, and 8 1,1-ADEQUATE correlations. The substructure analysis was applied to the 33 structures and identified 10 different substructures (Figure 4). This result is in accordance with a substructure analysis carried out by hand.[7b] The substructure analysis can be combined with the carbon chemical shift calculation (see Figure 1). The carbon chemical shifts for all substructure families of **1** are calculated and used for ranking (see Figure 5). Two substructures (**1**-S1 and **1**-S2) are clearly favored over the others. Substructures with a small

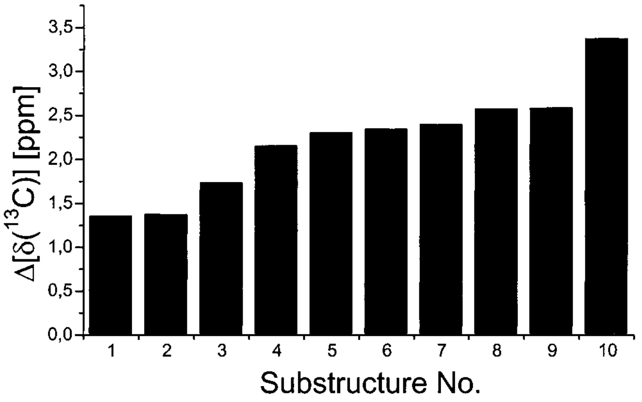

Figure 5. Results of the $\Delta[\delta(^{13}C)]$ calculation for all substructures generated for oroidin (**1**).

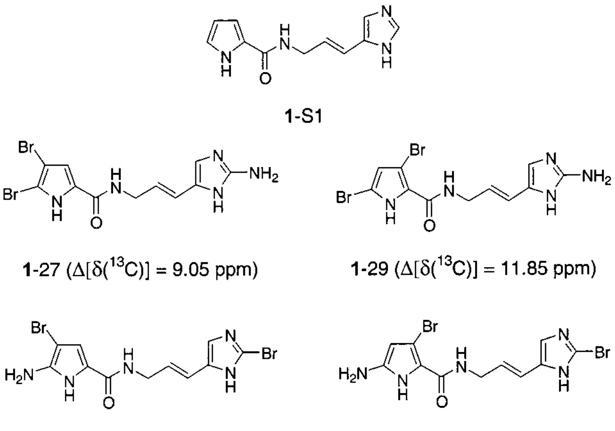
1-S1
1-27 (Δ[δ(13C)] = 9.05 ppm)
1-29 (Δ[δ(13C)] = 11.85 ppm)
1-30 (Δ[δ(13C)] = 18.40 ppm)
1-32 (Δ[δ(13C)] = 13.08 ppm)

Figure 6. Four structures of the best substructure family of oroidin (**1**).

deviation of their ^{13}C NMR spectrum with respect to the experimental spectrum have a high probability to be a part of the correct structure, since statistical errors in the chemical shift deviation are averaged out combining a large number of structures. If a multiple determination of a property value is possible, it is a known fact from statistical analysis that the precision and the accuracy of the prediction increase. In the described approach this fact leads to small deviations in the $\delta(^{13}C)$ prediction for substructures which are part of the correct solution. Since in a substructure several molecular structures are combined, $\delta(^{13}C)$ becomes the average value of $\delta(^{13}C)$ calculated for the individual molecules and tends to approach the experimental value. The two substructures differ in the connection of the pyrrole with the other part of molecule. In **1**-S1 the pyrrole is connected to the carbonyl carbon of the amide, whereas in **1**-S2 it is connected to the imidazole. Both could be distinguished because the carbamic acid bromides are not stable and the urea derivatives can be excluded by their $\delta(^{13}C)$.

The final substructure family (**1**-S1) consists of four structures (**1**-27, **1**-29, **1**-30, and **1**-32; see Figure 6). The structural proposals are numbered in sequence as generated by COCON. Structural proposals **1**-30 and **1**-32 which contain aminopyrrole and bromoimidazole substructures can be neglected because of the larger $\Delta[\delta(^{13}C)]$ and of different $\delta(^{15}N)$ in comparison to **1**-27 and **1**-29. The distinction of the 3,5-dibromopyrrole (**1**-29) versus the 4,5-dibromopyrrole (**1**-27) is possible by comparison of $\delta(^{13}C)$ of C-2, C-3, C-4, and C-5. The correct structure of oroidin (**1**, **1**-27) shows the lowest ^{13}C chemical shift deviation in this family (**1**-S1) and the second lowest of all 33 structural proposals. The absolute ^{13}C chemical shift deviations are rather high for this particular ensemble (from 7.8 to 20.5 ppm). However, only the relative information is of interest for this analysis. The relative large absolute $\delta(^{13}C)$ deviation has only a minor influence on the result.

Table 2. $\delta(^{13}C)$ Deviations for the 10 Best Structural Proposals of Manzacidin A (**2**) and 5-Deoxyenterocin (**3**)

manzacidin A (**2**)		5-deoxyenterocin (**3**)	
no.	$\Delta[\delta(^{13}C)]$, ppm	no.	$\Delta[\delta(^{13}C)]$, ppm
1	5.2	1	3.7
2	7.2	2	8.5
3	7.4	3	9.1
4	10.9	4	9.4
5	11.8	5	10.1
6	11.8	6	10.5
7	12.0	7	10.7
8	12.2	8	11.6
9	12.3	9	11.7
10	12.4	10	11.8

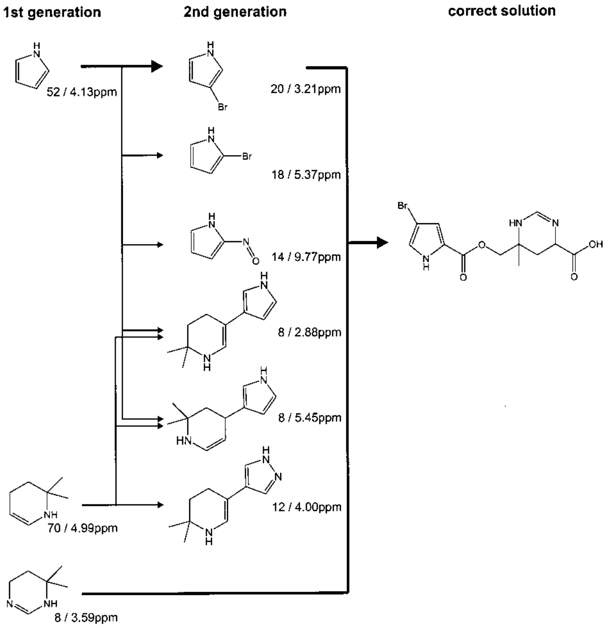

Figure 7. Substructure analysis of manzacidin A (**2**) allowing only two atoms to not be part of a ring in a substructure.

For the experimental data set of manzacidin A (**2**) including 6 $^{1}H,^{1}H$ COSY and 18 $^{1}H,^{13}C$ HMBC correlations, COCON generated 190 structural proposals. The results of the ^{13}C chemical shift calculation for the best 10 structures are given in Table 2. A part of the generated substructure tree of the manzacidin A (**2**) data set including all 190 structures is shown in Figure 7. The requirements for the substructures are (a) the minimum number of atoms per substructure is two, (b) the minimum number of molecules per substructure is eight, and (c) the substructures contain not more than two atoms that are not part of a ring system. The substructure analysis identified all different ring systems present in the ensemble. The chemical shift deviations for both ring systems (pyrrole and tetrahydropyrimidine) of manzacidin A (**2**) are smaller than those for other possibilities. The pyrrole subunit is found to be a part of 52 molecules, and the next generation of substructures is given here. The 3-bromopyrrole subunit

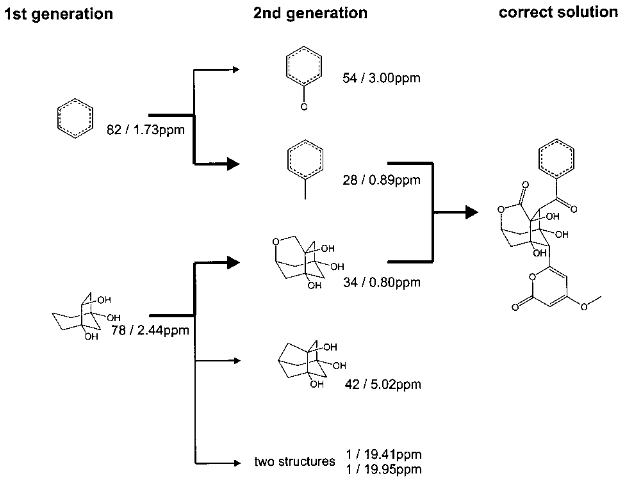

Figure 8. Substructure analysis of 5-deoxyenterocin (**3**) with at least six atoms and only one "non ring atom" per ring in a substructure.

of manzacidin A (**2**) is clearly preferred by its ^{13}C chemical shift deviation. Note that the sum of structures combined in two substructures in a subtree can be larger than the number of structures given at the root, since often two or even more substructures of a tree are present in one structure at the same time.

COCON proposed 82 structures for the experimental data set of 5-deoxyenterocin (**3**), which consists of 4 ^{1}H,^{1}H COSY (plus fixed phenyl ring) and 52 ^{1}H,^{13}C HMBC correlations. The results of the ^{13}C chemical shift calculation for the best 10 structures are given in Table 2. The substructure analysis of the 5-deoxyenterocin (**3**) data set presented as a substructure tree is shown in Figure 8, allowing only one "non ring atom" per ring system. The first two generations of substructures are given, and the substructures contained in the correct structure are indicated by bold bonds. The phenyl ring is found in all 82 structures, since it was fixed. However, two major groups can be found: 54 phenols and 28 carbon substituted benzenes. The second group is clearly favored by the chemical shift deviation (0.9 versus 3.0 ppm) and is also part of the correct proposal. The bicyclic system is found to be part of 78 out of the 82 structures. Again, two major groups were obtained in the next step introducing an additional bridge (tricyclic systems), one with and one without an oxygen. The oxygen-bridged substructure (oxymethylene) is favored by the lower chemical shift deviation in comparison to the methylene (0.80 versus 5.02 ppm) and is part of the correct solution.

In contrast to molecules **1**−**3** ascididemin (**4**) is more underdetermined with respect to the NMR correlation data set. To get some idea about the underdetermination of this system, a theoretical data set for **4** was generated including 14 ^{1}H,^{1}H COSY and 35 ^{1}H,^{13}C HMBC correlations. With this data COCON generated 28 672 structural proposals which show the requirement of C,C correlations or a fast method to analyze all structural proposals. The ^{13}C chemical shifts deviations between the experimental and the theoretical values were calculated for all 28 672 structures (see Figure 9). The correct structure of ascididemin (**4**) is ranked as 25th, which is within the first 0.1% of all structural proposals! The distribution over the carbon chemical shift deviation is Gaussian type (see Figure 9). This was also observed for other examples, but further investigations have to be carried out to derive any systematics.

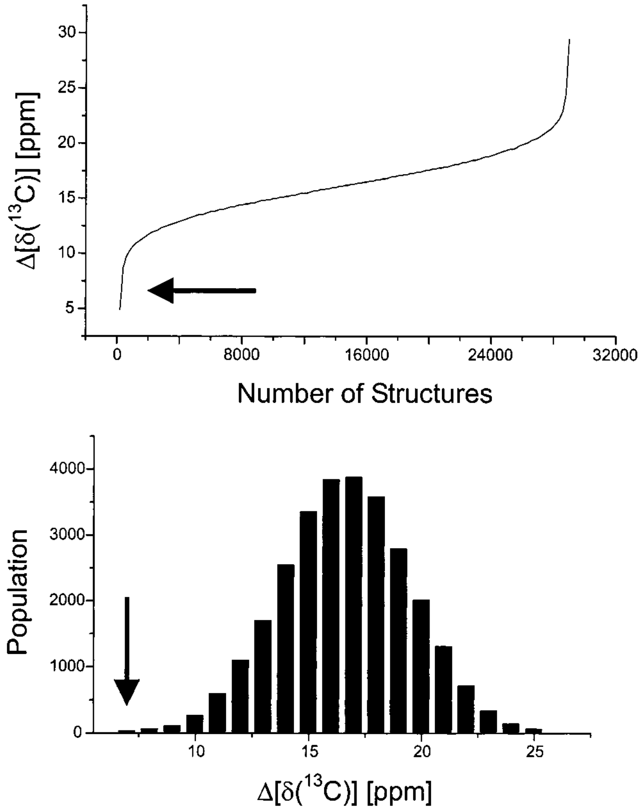

Figure 9. Results of the $\Delta[\delta(^{13}C)]$ calculation for all structural proposals of ascididemin (**4**) generated by COCON.

The substructure analysis cannot be applied to all generated structures due to computational requirements. Here, it is applied to the 300 structures with the lowest deviation of the calculated versus experimental $\delta(^{13}C)$ (about 1%). In contrast to examples **1**−**3**, the ^{13}C chemical shift deviations cannot be used as an argument for discrimination of substructures because these values are approximately the same for these structures (see Figure 9). However, substructure analysis can be used to investigate different ring systems present in this ensemble. Figure 10 shows the substructure analysis of **4** which results in 10 ring systems containing (a) at least 2 rings, (b) 10 atoms, and (c) that occur in at least 20 molecules. Again, the substructure of the correct solution has a rather small chemical shift deviation, but the differences from the others are not significant as mentioned before. However, the extraction of the basic ring systems in **4** give an overview about the set of structural proposals. Increasing the minimum number of required ring systems from two to three leads to 97 instead of 10 different ring systems.

For this example the complete way to the final structure will be discussed. Out of the best 60 structural proposals (approximately 0.2% of 28 672), there are only six nonstrained structures which do not violate Bredt's rule (see Figure 11). A further distinction is possible by taking $\delta(^{15}N)$ into account. Structures **4**-26112 (diazo), **4**-28613 (lactam), and **4**-28672 (nitroso) can be neglected using this argument. ^{1}H, ^{15}N HMBC correlations would be of help to distinguish between **4**-27927, **4**-28646, and **4**-28656. In structural proposal **4**-27927 there exists no nitrogen atom in the β-position to the carbonyl group. A correlation from the

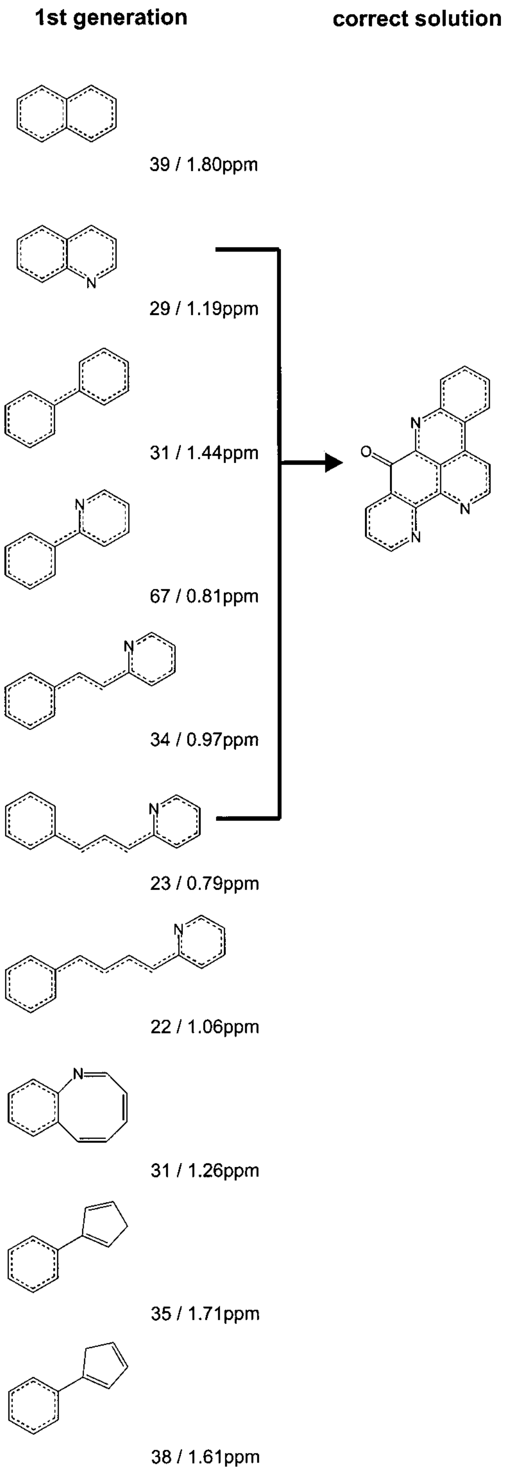

Figure 10. Substructure analysis of ascididemin (**4**). Substructures of the 300 structural proposals with lowest chemical shift deviation are given that combine (a) at least 20 molecules, (b) 10 atoms per molecule, and (c) 2 rings within a molecule.

phenyl ring to the nitrogen atom in the *β*-position to the carbonyl group is only possible for **4**-28646, which represents the correct constitution of ascididemin.

CONCLUSIONS

The widespread application of NMR-based structure generators such as COCON depends on the availbility of tools for the evaluation of the often large number of proposed constitutions. As long as every proposal would have to be analyzed by the operator, it can be expected that despite the availability of computer programs structures of new natural products will be published based on the insufficient criterion that they are in agreement with the spectral data.[18] The development of neural networks for the ^{13}C NMR based analysis of structural proposals closes a very important gap between a theoretical necessity and practical reality.

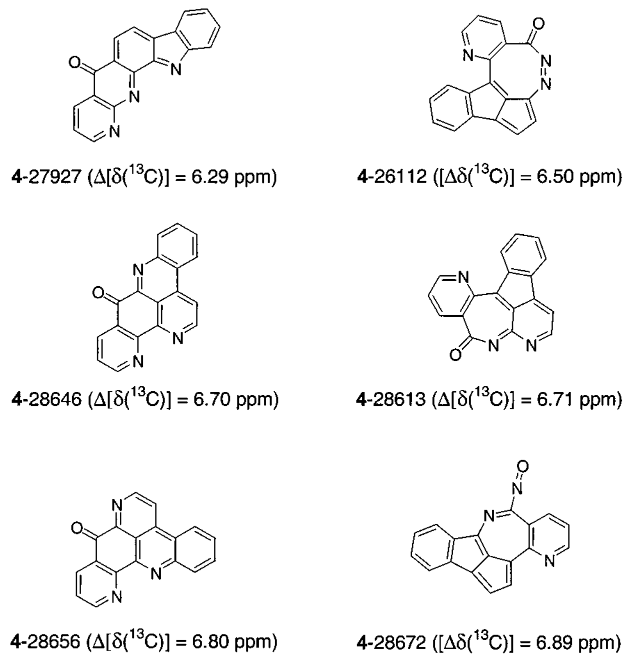
4-27927 (Δ[δ(13C)] = 6.29 ppm)
4-26112 ([Δδ(13C)] = 6.50 ppm)
4-28646 (Δ[δ(13C)] = 6.70 ppm)
4-28613 (Δ[δ(13C)] = 6.71 ppm)
4-28656 (Δ[δ(13C)] = 6.80 ppm)
4-28672 ([Δδ(13C)] = 6.89 ppm)

Figure 11. The six nonstrained structural proposals for ascididemin (**4**) out of the first 0.2% of 28 672 structures.

The usage of ^{13}C chemical shifts and substructure analysis allows analysis of structural proposals calculated by COCON. The short calculation times to obtain $\delta(^{13}C)$ by a neural network and the usage of a substructure analysis allow a structure elucidation with less correlation data from two-dimensional NMR spectra. This method combines advantages of both database and incremental prediction of chemical shift: it is accurate and fast. Since neural networks are able to interpolate, they can be applied to all kind of different organic substructures. However, a slightly lower accuracy can be expected for marine natural products because only a small number is incorporated in the underlying data set for training the neural networks. However, the absolute values of the chemical shift deviation are not important for this approach.[1] Only the deviation relative to the experiment is of interest. Therefore, the method does not essentially suffer from large absolute deviations since the structural proposals were generated from the same experimental data set.

The presented method is an alternative approach to obtain an almost complete correlation data set (including $^{1}H,^{15}N$ HMBC and $^{13}C,^{13}C$ correlation data) for an underdetermined structure. The number of structures that have to undergo a further analysis to obtain the correct result can be safely decreased to about 1% of the original number of structures for large ensembles without a significant risk of losing the correct proposal. This approach is independent of the structure generator COCON and can therefore also be used in combination with other structure generators. However, a combination of this approach with COCON is an essential step toward an automatic structure elucidation of organic compounds.

ACKNOWLEDGMENT

This research was supported by the Deutsche Forschungsgemeinschaft (Ko1314/3-3 and 3-4). J.M. is supported by a Kekulé stipend of the Fonds der Chemischen Industrie. We thank Christian Gödde for testing the substructure analysis extensively.

REFERENCES AND NOTES

(1) Köck, M.; Junker, J.; Maier, W.; Will, M.; Lindel, T. *Eur. J. Org. Chem.* **1999**, 579−586.

(2) Maier, W. In *Computer-Enhanced Analytical Spectroscopy*; Wilkens, C. L., Ed.; Plenum Press: New York, London, 1993; Vol. 4, p 37−55.

(3) Meiler, J.; Will, M.; Meusinger, R. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169−1176.

(4) (a) Bienfait, B. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 890−898. (b) Sharma, A. K.; Sheikh, S.; Pelczer, I.; Levy, G. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1130−1139. (c) Meiler, J.; Meusinger, R. In *Software*−*Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, pp 259−263. (d) Meusinger, R.; Moros, R. In *Software*−*Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, pp 209−216. (e) Montanarella, L.; Bassani, M. R.; Breas, O. *Rapid Commun. Mass Spectrom.* **1995**, *9*, 1589−1593. (f) Isu, Y.; Nagashima, U.; Aoyama, T.; Hosoya, H. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 286−293. (g) Michon, L.; Hanquet, B.; Diawara, B.; Martin, D.; Planche, J.-P. *Energy Fuels* **1997**, *11*, 1188−1193. (h) Rodrigues, G. d. V.; Campos, I. P. d. A.; Emerenciano, V. d. P. *Spectroscopy* **1997**, *13*, 191−200. (i) Zimmerman, D. E.; Kulikowski, C. A.; Huang, Y. P.; Feng, W. Q.; Tashiro, M.; Shimotakahara, S.; Chien, C. Y.; Powers, R.; Montelione, G. T. *J. Mol. Biol.* **1997**, *269*, 592−610. (j) Amendolia, S. R.; Doppiu, A.; Ganadu, M. L.; Lubinu, G. *Anal. Chem.* **1998**, *70*, 1249−1254. (k) Kaartinen, J.; Mierisova, S.; Oja, J. M. E.; Usenius, J.-P.; Kauppinen, R. A.; Hiltunen, Y. *J. Magn. Reson.* **1998**, *134*, 176−179.

(5) (a) Kvasnicka, V.; Sklenak, S.; Pospichal, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747. (b) Doucet, J.-P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324. (c) Sklenak, S.; Kvasnicka, V.; Pospichal, J. *Chem. Pap.* **1994**, *48*, 135−140. (d) Clouser, D. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168−172. (e) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.-P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653. (f) Azzouzi, S. R. E.; Fan, B. T.; Panaye, A.; Doucet, J.-P. *Org. React.* **1997**, *31*, 3−20. (g) Meiler, J.; Meusinger, R.; Will, M. *Monatsh. Chem.* **1999**, *130*, 1089−1095.

(6) (a) Shemtulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871. (b) Ozawa, K.; Yasuda, T.; Fujita, S. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 688−695. (c) Robinson, D. D.; Barlow, T. W.; Richards, W. G. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939−942. (d) Wang, T.; Zhou, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 828−834. (e) Klein, D. J.; Gutman, I. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 534−536.

(7) (a) Lindel, T.; Junker, J.; Köck, M. *J. Mol. Model.* **1997**, *3*, 364−368. (b) Lindel, T.; Junker, J.; Köck, M. *Eur. J. Org. Chem.* **1999**, 573−577. (c) Junker, J.; Maier, W.; Lindel, T.; Köck, M. *Org. Lett.* **1999**, *1*, 737−740. (d) Köck, M.; Junker, J.; Lindel, T. *Org. Lett.* **1999**, *1*, 2041−2044.

(8) (a) Forenza, S.; Minale, L.; Riccio, R. E. F. *J. Chem. Soc., Chem. Commun.* **1971**, 1129−1130. (b) Garcia, E. E.; Benjamin, L. E.; Fryer, R. I. *J. Chem. Soc., Chem. Commun.* **1973**, 78−79. (c) Walker, R. P.; Faulkner, D. J.; van Engen, D.; Clardy, J. *J. Am. Chem. Soc.* **1981**, *103*, 6772−6773.

(9) Kobayashi, J.; Kanda, F.; Ishibashi, M.; Shigemori, H. *J. Org. Chem.* **1991**, *56*, 4574−4576.

(10) Kang, H.; Jensen, P. R.; Fenical, W. *J. Org. Chem.* **1996**, *61*, 1543−1546.

(11) Kobayashi, J.; Cheng, J.; Nakamura, H.; Ohizumi, Y.; Hirata, Y.; Sasaki, T.; Ohta, T.; Nozoe, S. *Tetrahedron Lett.* **1988**, *29*, 1177−1180.

(12) (a) Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 151−162. (b) Bremser, W. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(13) *SpecInfo database*; Chemical Concepts: Karlsruhe, 2001.

(14) (a) Robien, W. *Monatsh. Chem.* **1983**, *114*, 365. (b) Robien, W. *Nachr. Chem. Technol. Lab.* **1998**, *46*, 74−77.

(15) (a) Schweitzer, R. C.; Small, G. W. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 46−53. (b) Schweitzer, R. C.; Small, G. W. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 310−322.

(16) (a) Clerc, J.-T.; Sommerauer, H. *Anal. Chim. Acta* **1977**, *95*, 33−40. (b) Ewing, D. F.; Abe, K. *Org. Magn. Reson.* **1979**, *12*, 499−524. (c) Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A. *Carbon-13 NMR Spectral Data*; Verlag Chemie: Weinheim, 1981. (d) Hearmon, R. A. *Magn. Reson. Chem.* **1986**, *24*, 995−998. (e) Fürst, A.; Pretsch, E. *Anal. Chim. Acta* **1990**, *229*, 17−25. (f) Thomas, S.; Ströhl, D.; Kleinpeter, E. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 725−729.

(17) Meiler, J. www.jens-meiler.de, 2001.

(18) Faulkner, D. J. In *Marine Biotechnology*; Attaway, D. H., Zaborsky, O. R., Eds.; Plenum Press: New York, 1993; Vol. 1, pp 459−474.

CI010294N