Fast Determination of ^{13}C NMR Chemical Shifts Using Artificial Neural Networks

J. Meiler,*,† R. Meusinger,‡ and M. Will§

Institute of Organic Chemistry, Marie - Curie - Strasse 11, University of Frankfurt, D-60439 Frankfurt, Germany, Institute of Organic Chemistry, University of Mainz, D-55099 Mainz, Germany, and BASF AG Ludwigshafen, D-67056 Ludwigshafen, Germany

Received March 15, 2000

Nine different artificial neural networks were trained with the spherically encoded chemical environments of more than 500 000 carbon atoms to predict their ^{13}C NMR chemical shifts. Based on these results the PC-program "C_shift" was developed which allows the calculation of the ^{13}C NMR spectra of any proposed molecular structure consisting of the covalently bonded elements C, H, N, O, P, S and the halogens. Results were obtained with a mean deviation as low as 1.8 ppm; this accuracy is equivalent to a determination on the basis of a large database but, in a time as short as known from increment calculations, was demonstrated exemplary using the natural agent epothilone A. The artificial neural networks allow simultaneously a precise and fast prediction of a large number of ^{13}C NMR spectra, as needed for high throughput NMR and screening of a substance or spectra libraries.

INTRODUCTION

NMR spectroscopy is undoubtedly one of the most important methods used for structure determination of chemical compounds. In recent years the power of NMR methods and the sophistication of spectrometers increased clearly. This was achieved by a number of new techniques.[1] Only a few of them should be named here. The measurement time was decreased drastically by pulsed field gradients, double or single quantum coherence methods, and finally by the so-called "tubeless NMR". This is the fitting of conventional high-resolution NMR spectrometers with flow-probes or special micro sample probes. Shorter NMR measuring times are required above all by the high throughput methods developed in combinatorial chemistry. With increasing amounts of spectral data available a new bottleneck has emerged: data analysis. Precise and fast computer programs are necessary to enhance the productivity here. Munk gave recently a vivid presentation of the evolution of computer enhanced structure elucidation exemplary by the structure determination of the antibiotic actinobolin.[2] In the 1960s the computer assisted elucidation of unknown structures required several man years using the structure generator ASSEMBLE. Forty years later, with both, more sophisticated NMR spectroscopic methods and computer software, the time required to determine the structure has been reduced to several days (time for data collection included). Now the program SESAMI generated four candidate structures in 5 min CPU time using only the available 1D and 2D NMR data. Lindel et al. also use both the NMR spectroscopic detectable connections between nuclei and their chemical shifts,[3] in their program COCON (*co*nstitutions from *con*nectivities) which was developed for the generation of all possible constitutions for complex natural products. The efforts which are spent for the development of efficient structure elucidation programs shall be presented here by two other current examples. CISOC-SES[4] is a computer assisted expert system that utilizes 1D and 2D NMR data. Recently the NMR assignment of a biologically active triterpenoid was shown by Peng et al.[5] With the program LSD (*L*ogic for *S*tructure *D*etermination) Nuzillard demonstrated impressively the potential of systematic structure elucidation of small molecules combining modern NMR spectroscopy with artificial intelligence at the example of gibberellic acid.[6] However, in most practical cases an elucidation of a completely unknown structure is not required. The more common type of structure determination is the structure verification. In this case, enough information is available perhaps on the basis of well-known synthetic reaction paths to propose a probable structure. The structure information which is achieved via the chemical shift is usually sufficient here.

NMR Chemical Shift Prediction. Atomic nuclei of one isotope located within one molecule in different chemical environments are shielded differently by their electron cloud. As a result, different resonance frequencies are observed during an NMR experiment exciting these isotopes. If these frequencies are measured as differences to the resonance frequency of an inner standard, they are designated as "chemical shifts". The chemical shift value combines two advantages for structural analysis. It is an easily obtainable spectral parameter, and its dependence on chemical structure is well-known.[7] The chemical shift of a carbon is, in addition to its state of hybridization, mainly influenced by the kind and number of the bond atoms and by their distances to the observed carbon. The chemical shift of a carbon atom can be influenced by another atom in two different ways: electron interaction over covalent bonds or through space. In solution the second effect appears possibly as a "solvent effect". However, electron interaction through space is only important if the distance between the observed and influencing atom is small. It has to be considered specially during

* Corresponding author phone: ++49 69 798 29 798; fax: ++49 69 798 29 128; e-mail: mj@org.chemie.uni-frankfurt.de.
† University of Frankfurt.
‡University of Mainz.
§ BASF AG Ludwigshafen.

10.1021/ci000021c CCC: $19.00
Published on Web 08/25/2000

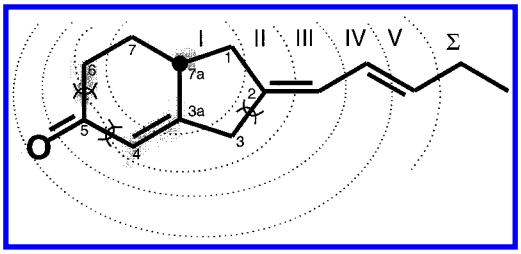

Figure 1. Schematic representation of the spherical division of the chemical environment of an carbon atom in a molecule. The carbon 7a of the substituted 1,2,3,6,7,7a-hexahydroinden-5-one is selected as focus (●). Five spheres starting from this focus are shown (I to V). The beyond environment is summarized in a "sum-sphere" (Σ). The π-contact areas of the focus carbon were marked with a gray background and the ring closure elements with)((see text).

stereochemical analyses. The stronger effect is transmitted via the orbitals following covalent bonds. It depends on the number and the σ- or π-type of the connecting bonds. Focusing on a randomly chosen atom in a molecule, one can consider all other atoms of this molecule as members of spheres. These spheres surround the focused atom, and their number is identical to the number of bonds between the focused atom and the atoms combined in this sphere. In Figure 1 the subdivision of a molecule into five different spheres (Roman numerals) with respect to one carbon atom (focus) is shown. In principle this procedure was already described in 1973 by Bremser with his well-known HOSE code (*h*ierarchically *o*rdered *s*pherical description of *e*nvironment).[8] The influence of the substituents on the chemical shift of the focused carbon normally decreases with an increasing number of bonds between them and therefore also with the increase of the sphere. By σ-bonded atoms the electrons are kept in the molecular orbitals located between two atoms. Therefore the influence decreases here fast with increasing sphere. By far different is the situation in the case of atoms bonded over a conjugated π-electron system. The electron density distribution can be influenced over more than one bond, and larger effects are induced over four, five, or even more bonds.

A further advantage of the NMR chemical shift is the availability of huge amounts of experimental data. Several databases exist today containing hundreds of thousands of chemical shift values in particular for ^{13}C nuclei and the appending information about the chemical environment of these individual carbon. These data are an excellent basis for computer assisted structure determination. Only some examples of electronically stored databases should be mentioned here: SPECINFO,[9] a further development of the ^{13}C spectral database created by Bremser et al.;[10] CSEARCH,[11] a database created by Robien in the beginning of the 1990s; WINDAT,[12] a database created by Trepalin and co-workers; and the CNMR[13] database which was developed few years ago by the ACD company.

Looking closer, ^{13}C NMR spectra databases are statistical tools to establish the relationships between NMR spectral parameters and the chemical environment of individual carbon atoms to propose either chemical structures or spectra. The ^{13}C chemical shift value is extremely suitable for this purpose because of its accuracy, reproducibility, and intelligible structure dependence. Using a large collection of representative molecules for structure determination two approaches are thinkable: starting from an experimentally determined chemical shift value a suitable chemical environment can be determined and starting from a chemical structure the relevant chemical shift can be estimated. Consequently, ^{13}C NMR spectra databases are suitable for three applications: (1) the prediction of NMR parameters for any molecular structure, (2) the verification of existing assignments, and (3) the determination of one or more possible molecular structures corresponding to a ^{13}C NMR spectrum. (This contains the simultaneous assignment of individual NMR signals to the respective carbon of a known structure.)

However, the results can only be offered with a statistical probability, depending on the quantity and quality of the available database entries. In addition to the time required for compiling the data, the greatest problem of database management is the accuracy and reliability of the represented data. In other words, the accuracy of the predicted chemical shifts, chemical environments, or structures cannot be more precise than the stored data. For this reason, careful examinations of shift assignments are just as important as the typographical error checking. Usually, some quality checking procedures are applied. In particular the assignments for quaternary carbons have been twisted in a number of cases, because an unambiguous experimental assignment was often not possible in the past.

Two further possibilities for estimating the assignment between chemical structures and chemical shifts besides database search should be mentioned here: the calculation of the chemical shift values by applications of empirical methods and the computation by quanta chemical procedures. However, quanta chemical computations, e.g. with the IGLO-method (*i*ndividual *g*auge for *l*ocalized *o*rbitals),[14] are relatively extensive. Meanwhile the prediction of ^{13}C NMR spectra is one of the most intensely studied applications of empirical modeling. This method is based on the assumption that the influence of different substituents on the chemical shift of an individual carbon atom can be defined simply by a set of constant values, the "increments". According to the chemical environment of a carbon atom all increments are added up. The chemical environment is defined by the kind and number of the neighboring atoms or atomic groups and by their distances to the considered carbon atom, in this case. The increments themselves were determined by multiple linear regression analysis using data sets of observed chemical shifts from structurally related compounds. The mean advantage of these increments is their simple application and the shortness of its computation. However, the increments are structure class dependent and available only for some substance classes and/or structure groups,[15] e.g. for alkanes,[16] alkenes,[17] substituted benzenes,[18] naphthalenes, and pyridines.[19] Different PC programs were developed in the past which allow the computation of more complex structures, for example the programs SPECTOOL[20] created by Pretzsch et al. and CSPEC2[21] developed by Cheng and Kasehagen. However, the increments employed in these programs were obtained from individual representatives of single substance classes too, and one has to be careful during their application. Possible interactions between several increments are often not considered by these approaches. This was recently shown for the aromatic carbons in substituted benzenes.[22] In such a way, the structure analyst must

currently decide between two possibilities: Either for a precise prediction requiring a long time or for the rapid available information afflicted with a larger error. For this reason, several attempts were taken in the past to describe the association between NMR chemical shifts and chemical structure more precisely. Different nonlinear numerical and statistical techniques, such as principal component analysis and artificial neural networks, were used.

Neural Networks. The main advantage of neural networks compared to other methods is their greater capacity to extract general information from a training data set and apply it on presented new data. Neural networks appeared in chemistry at the beginning of the 1990s.[23] The first publications dealing with the determination of ^{13}C NMR chemical shift values using neural networks were also published at this time. Kvasnicka et al. determined the chemical shifts of carbons in monosubstituted benzenes,[24−26] and Doucet and coworkers predicted the shifts of C5 to C9 alkanes.[27] Anker and Jurs[28] trained a three layer neural network with a data set of 391 steroid carbon atoms using the back-propagation learning algorithm. The applied network architecture consisted of 13 input units corresponding to calculated atom-centered descriptors, 40 hidden neurons, and 116 output neurons corresponding to 0.5 ppm chemical shift increments in the range of 8.7−66.7 ppm. The examined results were superior to those achieved with linear regression techniques in every case. Further works followed, all of them with the common characteristic of ^{13}C NMR chemical shift prediction for a group or a class of substances with similar chemical structures, e.g. for alkanes,[29−31] cyclohexanes,[32] alkenes,[33,34] substituted naphthalenes,[35] trisaccharides,[36] dibenzofuranes,[37] ribonucleosides,[38] and substituted benzenes.[22] On the assumption that the influences of substituents on an observed carbon are only similar within an individual substance class, the chemical shift values computed with such a neural network cannot been generalized as desired. Consequently, this is not different from the respective increment systems, certainly more precise but limited in their applicability though. Therefore a fast method for computation of most organic molecular structures is desirable similar as by the use of large databases.

Another approach combining these advantages might be the use of genetic algorithms.[39] However it was not tested yet for this purpose and will not be discussed in this paper.

Molecular Structure Descriptors. In order to calculate the chemical shift values of carbon atoms with neural networks it is necessary to describe the atom types and the chemical environments of all atoms numerically. An optimum must be found for the number of the applied descriptors. On the one hand, the number must be as small as possible for computational reasons; on the other hand, the descriptors must feature clearly the differences in molecular structures. In total 28 descriptors were determined for 28 different atom types and are summarized in Table 1. Additionally, the numbers of representatives of these atom types found in a data base of about 40 000 molecules[9] with up to 100 heavy atoms are indicated. Atom types are derived from element number, hybridization state, and number of bonded hydrogen atoms. As given in Table 1, nine different atom types were defined for the 526 565 carbon atoms which were available for the prediction of ^{13}C NMR chemical shifts in total. In some cases, different atoms were summarized into one type if they only differ in the number of bonded hydrogen atoms. This was necessary since the number of the representatives would have been too small, for example in the case of olefinic carbons with one or two hydrogen atoms (type 6), for sp^2 hybridized nitrogen (type 13), and for sp^3 hybridized phosphorus (type 20) and sulfur (type 22). Two further descriptors were introduced in order to obtain a complete description. One descriptor holds the number of all hydrogen atoms located in the individual sphere (type 29). In order to consider the influence on the ^{13}C NMR chemical shift caused by the formation of rings, a second sum descriptor holds the number of ring closures (type 30). Two different possibilities have to be distinguished: If the ring is closed within one sphere the descriptor increased by two. That is the case in odd-numbered rings (3, 5, ..., $2n + 1$ atoms with $n \epsilon N$), whereas in even-numbered rings (4, 6, ..., $2n$ atoms with $n \epsilon N$) the ring is closed over two spheres. In this case the descriptor increased by one for both spheres affected. In the example visualized in Figure 1, the five-membered ring is closed between carbons C-2 and C-3 within the sphere II. In contrast the six-membered ring must be closed between carbons C-4 and C-6, both in sphere II, and carbon C-5 in sphere III.

The chemical environments of the carbons are described by sorting the atoms in spheres as given in Figure 1 and counting the occurrence of every atom type in each sphere. As shown in Figure 1, five spheres (I to V) are formed for every carbon. All atoms in further distances are projected in an additional "sum sphere" (Σ). Consequently, 30 numbers are necessary to describe one sphere, and 180 numbers are necessary for the complete description of the environment of an individual carbon. However, this description is only based on number, kind, and distances of the substituents, which is not sufficient. In addition to this nonspecific enumeration, the descriptors are extended by a so-called "π-contact area", in order to take the special importance of conjugated π-electronic systems into consideration. Two atoms are in "π-contact" if a conjugated π-electronic system exists from at least one neighbor of the first atom to one neighbor of the other atom. The two atoms themselves are not taken into consideration for this parameter since their belonging to the conjugated system is given by their atom type. Therefore, by description of the environment of carbon C-7a in Figure 1 (focus), the double bonds in the side-chain must not be considered as π-contact, since they do not belong to a conjugated π-electronic system that includes one neighbor of the focused carbon. Otherwise, two sp^3-hybridized carbon atoms have to be considered here. The carbons C-3 and C-6 are neighbors of the conjugated π-electronic system (C3a═C4-C5═O) just as the observed carbon C-7a. In Figure 1 the entire π-contact area of the carbon C-7a is shown as shaded region. The molecular structure descriptors are extended by a second set of 30 numbers for each sphere using the atom types and sum parameters described above in Table 1. But in contrast to the general count in the first set only the atoms being in π-contact with the focused carbon were considered now. Consequently, for each sphere two sets of descriptors result for the encoding of the environment. This is shown in Figure 2. The structure of the descriptors for the calculation of the chemical shift of an individual carbon is represented schematically for the first sphere on the top and for all spheres in the middle of Figure 2.

Table 1: Atom Types (1−28) and "Sum-Types" (29−30) Defined for Describing Atom Environments and Their Frequencies in the Data Set

ID	atom type	frequency of atoms of this type
1	⟩C⟨	19 527
2	⟩CH-	49 556
3	-CH_2-	116 175
4	-CH_3	73 724
5	=C⟨	50 711
6	=CH-/=CH_2	27 556
7	≡C-/≡CH/=C=	3793
8	) ⟩C- (aryl)	68 416
9	) ⟩CH (aryl)	117 107
10	⟩N-	9876
11	-NH-	9115
12	-NH_2	3521
13	=N-/=NH	7427
14	≡N	2053
15	-NO_2	2688
16	) ⟩N (aryl)	3743
17	-O-	31 641
18	-OH	20 626
19	=O	39 259
20	⟩P-/-PH-/-PH_2	383
21	⟩PO-	1053
22	-S-/-SH	4146
23	=S	1214
24	⟩SO_2	1789
25	-F	3613
26	-Cl	9585
27	-Br	2718
28	-I	603
29	sum of all hydrogen atoms bond in this sphere	
30	sum of all ring closures in this sphere	

EXPERIMENTAL SECTION

A total of 40 000 molecules were available for the calculations. The 526 565 carbons with well-known ^{13}C NMR chemical shifts and molecular structure assignments[9] were distinguished the nine mentioned atom types (Table 2). All chemical shifts were estimated in deuterated chloroform ($CDCl_3$) or in carbontetrachloride (CCl_4) and refer to tetramethylsilane (TMS) as internal standard. With this, solvent effects were excluded as far as possible. Stereochemical information was not available for the given molecules. The environment of every individual carbon atom was encoded with the descriptors from Table 1. Three hundred sixty descriptors were used as input vector for the neural networks, whereas the individual ^{13}C NMR chemical shift value represents the output. For each of the nine atom types representing a carbon atom an individual neural network was constructed and trained using back-propagation of errors.[40] The amount of available molecules was randomly subdivided into three sets. Ninety percent of data was used for the training of the neural networks. During the training process the percentage of data was increased stepwise up to 90% maximum. A second data set contained 7% of the available data for monitoring. The training and monitoring data sets were used simultaneously. To avoid "overtraining", the iterative training process was stopped if the deviation for the monitoring data set increased again. The hyperbolic tangent (tanh) was found to give best results as a transfer function. The number of hidden neurons was optimized to give the lowest error for the monitoring set of data. Finally, the third data set of 3% randomly selected molecules was used as an independent set for testing the trained networks.

The PC program C_shift[41] which included the trained neural networks and a simply manageable structure-editor for the structure input was written in C++ for Windows 95/98/NT.

RESULTS AND DISCUSSION

In Figure 2 the architecture of neural networks used here is shown schematically. While the number of input and output units was fixed, the number of the hidden neurons had to be determined experimentally. The best results were achieved with a number of 5−20 hidden neurons, depending on the carbon atom type. In Table 2 the descriptions are compiled for the nine different neural networks. The results achieved with these nets for the training and testing data sets are also given in the form of statistical information, respectively. It is shown that a mean deviation of 1.79 ppm and a standard deviation of 2.10 ppm resulted for the more than 15 000 carbons in the test data sets. The accuracy of these calculations is much better in comparison to systems based on fixed increments, especially if complex structures are investigated. In Figure 3 the correlation between the computed and the experimental chemical shift values are shown for all nine carbon atom types obtained from the test data sets. The appropriate correlation coefficients are given in Table 2.

On closer examination of these results it is striking that the largest deviations were obtained for the sp-hybridized carbons (standard deviation of 3.8 ppm for type 7) and for carbons with sp^2 hybridization (standard deviations of 3.7 and 3.6 ppm for types 5 and 6), respectively. For triple bonded carbons the number of examples was relatively small. That causes the higher uncertainty in this calculation.

Furthermore one has to note that all effects influencing the chemical shifts through space were not considered here. Therefore comparable carbons in different stereoisomers do not become distinguishable through their different spatial environments. It is well-known that the different arrangements of substituents either in the *E*- or in the *Z*-direction can influence the chemical shift of the olefinic carbons up to 5 ppm. Influences in this order of magnitude were also observed in different configurational and conformational isomers caused by the dissimilar spatial arrangements of their substituents. Furthermore, one observes an increase of the uncertainty for sp^3 hybridized carbons with an increasing number of non-hydrogen substituents. Here, the standard deviation increases from 1.3 ppm for methyl group carbons up to 3.4 ppm for the quaternary carbons (Table 2). This is expected, since the chemical shift of a methyl group with three fixed substituents in the first sphere is much easier to determine than the chemical shift of a quaternary carbon where four different substituents can interact in the first sphere already.

As already mentioned, the influences of the substituents on the chemical shifts depend on their distances to the observed carbon and on its affiliation to a conjugated π-system. This knowledge can also be obtained analyzing the weights of the trained neural networks. Figure 4 shows the sensitivities of the input units of three different neural networks, subdivided in the six spheres and the types of covalent bonds. These values were determined for every input unit, by variation of their inputs, while all other inputs stay constant at zero. The sensitivity of the selected input is given by the range detected at the output. In Figure 4 the sums of

Table 2: Number of Hidden Neurons, Frequencies of Atom Types (1−9), Correlation Coefficients, Standard Deviations (in ppm), and Mean Deviations (in ppm) for the Training and Test Data Sets of the Nine Different Carbon Atom Types

			training and monitoring data				test data			
ID	atom type	no. hidden neurons	count	corr coeff	std [ppm]	m.d. [ppm]	count	corr coeff	std [ppm]	m.d. [ppm]
1	⟩C⟨	10	18 984	0.995	1.83	1.71	543	0.983	3.42	2.87
2	⟩CH-	10	48 032	0.984	2.37	2.44	1542	0.984	2.39	2.47
3	-CH_2-	20	112 639	0.982	1.85	1.51	3536	0.985	2.65	1.39
4	-CH_3	20	71 547	0.989	1.47	1.16	2177	0.988	1.30	1.01
5	=C⟨	10	49 139	0.989	2.71	2.71	1518	0.981	3.68	2.72
6	=CH-/=CH_2	10	26 691	0.959	3.78	3.28	865	0.966	3.60	3.46
7	≡C-/≡CH/=C=	5	3675	0.995	1.99	2.18	118	0.988	3.80	2.96
8	) ⟩C- (aryl)	20	66 433	0.982	1.88	2.02	1983	0.981	1.72	1.84
9	) ⟩CH (aryl)	20	113 655	0.971	1.57	1.13	3452	0.963	1.81	1.35
all			***510 795***	***0.981***	***1.97***	***1.75***	***15 716***	***0.979***	***2.10***	***1.79***

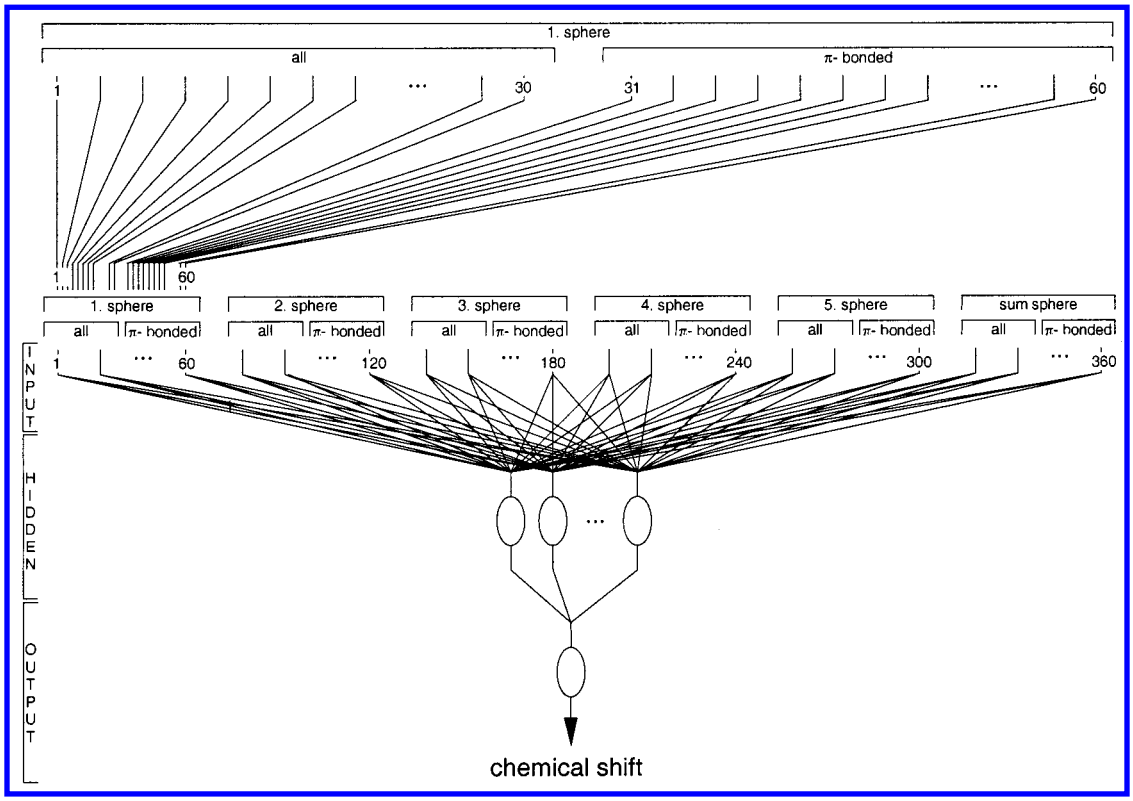

Figure 2. Schema of a three layer neural network for calculating chemical shifts. Sixty numbers divided in two sets result as input for the individual spheres of the environment, respectively. Thirty numbers were used for all substituents (all), and the second set of 30 numbers hold the count of the atoms in the π-contact areas (π). In total 360 input values were required for the five individual spheres and the additional sum sphere. The number of hidden neurons varied between 5 and 20. The single output neuron predicts the chemical shift of the observed carbon atom.

the sensitivities for 30 input neurons for every sphere and every bonding type is shown for three different carbon atom types, respectively. Methyl groups (Figure 4a) shown a fast decrease of the influence of substituents with increasing sphere number. For all atoms beyond the third sphere the influence is low at all. This is in accordance with the knowledge of the substituents induced shifts. In contrast to this the atoms with π-contact show distinctly smaller influences. A significant increase of the sensitivity against atoms with π-contact was observable for the double-substituted sp^2 hybridized carbons (Figure 4b). Since most of the conjugated systems do go beyond the second sphere, a strong decrease was observed after this sphere. Finally the influence of π-conjugated systems can be seen clearly in aromatic systems (Figure 4c). The sensitivity for π-bonded atoms is constantly high over the first three spheres in order to decrease slowly behind the third sphere. Only in the first sphere was observed a smaller influence in comparison to the sum of all substituents.

To test the capability of the method for the determination of ^{13}C NMR chemical shifts, the well-known epothilone A was chosen as an example (Figure 5). Whereas the structure and the NMR spectra of this natural cytotoxic agent is described in detail,[42] this molecule was not included into the database[9] and is therefore used for a comparative determination. Consequently, epothilone A was also not a part of the training or testing data set of the neural networks. In Table 3 all experimentally determined ^{13}C NMR chemical shifts[42] are listed as well as the values calculated by use of the neural networks[41] and the values predicted from a database after spherical coding of all carbons with the HOSE code. Experimental chemical shifts were only available for determination in DMSO. So, the agreements between experimental and computed values were not as precise as

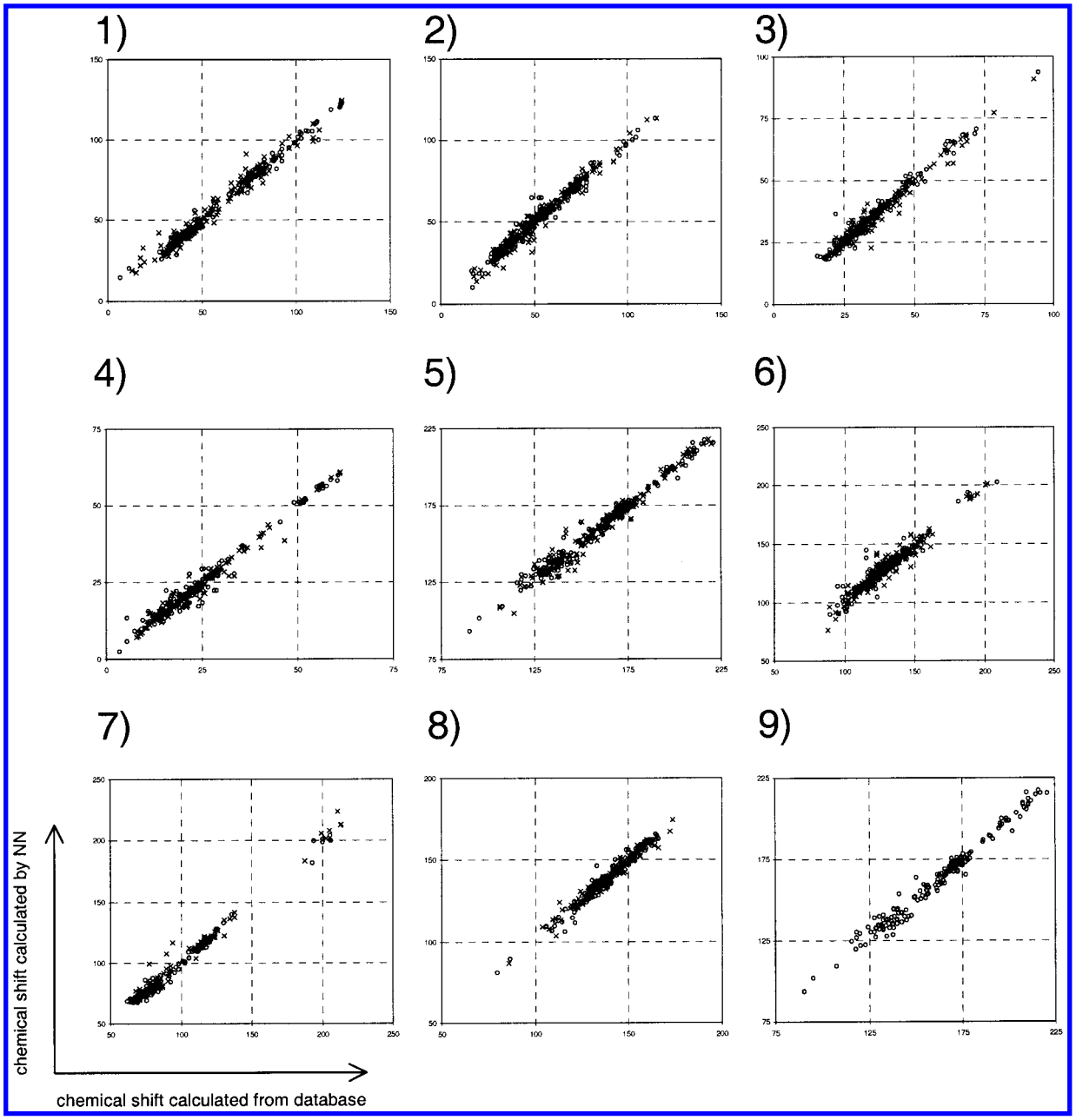

Figure 3. Correlation of chemical shifts of carbons from test data sets determined by use of the data base (x-axis) and by neural networks (y-axis). The results are shown for the nine different types of carbon atoms. The numbering of the diagrams is identical to the ID used in Tables 1 and 2. All data are given in ppm with respect to TMS. For the correlation coefficients and other statistical results look at Table 2.

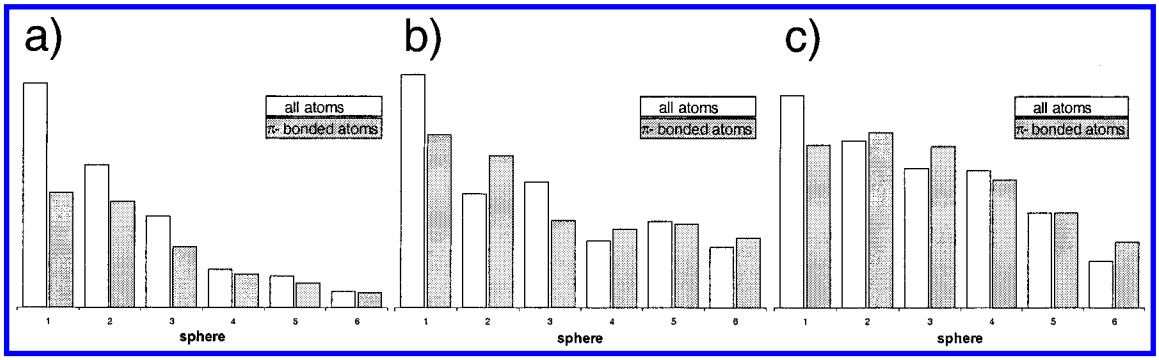

Figure 4. Sum of the sensitivity of the input-units depending on spheres and the types of the covalent bonds for the neural networks trained (a) for methyl carbons (atom type 4), (b) for olefinic carbons (atom type 5), and (c) for aromatic carbons (atom type 8).

expected. However, this restriction concerned all determination procedures uniformly. The standard deviations found here were 2.5 ppm for the neural networks and 2.4 ppm for the determination using the HOSE code description. There-

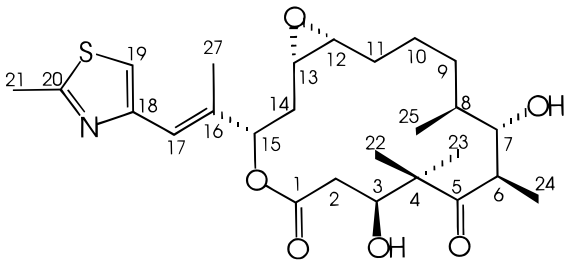

Figure 5. Structure of epothilone A.

Table 3: ^{13}C NMR Chemical Shifts (in ppm) of Epothilone A[a]

	chemical shift [ppm]			
atom ID (acc. Figure 5)	exptl	neural net	specinfo	spectool
1	170.2	173.3	172.2	172.0
2	38.3	36.6	38.9	36.1
3	70.9	73.7	72.8	73.4
4	53.0	52.6	52.8	53.6
5	216.9	216.9	215.9	217.0
6	45.2	45.3	42.3	44.6
7	75.7	75.4	73.5	75.3
8	35.3	33.5	35.0	34.3
9	29.5	32.3	27.7	30.7
10	23.3	24.8	23.7	24.0
11	26.6	30.9	29.9	31.6
12	56.4	61.1	57.9	56.6
13	54.3	58.0	53.6	50.9
14	31.9	35.1	31.0	33.1
15	76.2	76.4	74.7	78.7
16	137.1	137.8	138.7	143.0
17	118.9	122.2	114.2	120.6
18	151.8	159.5	149.2	142.5
19	117.5	116.4	123.1	118.7
20	164.0	169.5	165.7	165.9
21	18.6	18.9	18.9	16.2
22	22.4	21.6	20.6	14.6
23	20.6	21.6	20.6	14.6
24	16.5	15.7	14.6	8.4
25	18.6	15.8	13.2	13.9
27	14.0	16.7	16.6	10.7
mean dev:		2.2	1.9	2.9
std dev:		2.5	2.4	3.8
corr coeff:		0.9991	0.9991	0.9980

[a] The shifts were determined experimentally in DMSO and calculated by the neural network,[41] by the HOSE code based estimation performed with the Specinfo data-base,[9] and by increments (SpecTool).[20]

fore, the results are of a comparable quality. But the advantage of the neural network approach was its drastically shorter computation time. The result was computed 1000 times faster compared to the HOSE code based determination for which the entire database had to be searched. The chemical shifts determined with the increments show a clearly larger deviation (3.8 ppm), as expected. A more detailed look reveals the largest deviations for the adjacent carbons to the epozid (C-11 to C-14), the carbons included in the conjugated π-system of the five-membered ring (C-16 to C-20) and for the methyl groups (C-21 to C-27). Obviously all three methods have difficulties with the conjugated π-system in the heterocyclic ring. While the adjacent carbons to the epoxid were computed quite precisely with the database, relatively large deviations are observed here for the results obtained with the neural networks. This could be caused by stereochemical influences in the high-flexible alicyclic part of the molecule.

CONCLUSIONS

Artificial neural networks offer a fast and accurate possibility to calculate ^{13}C NMR chemical shifts of organic compounds. While the method has an essentially lower standard deviation than increment methods, the computation time is 1000 times faster than using comparable accurate database predictions of chemical shifts. This makes neural networks ideal for screening results of stucture generators or checking the entries of a database. If a large number of ^{13}C NMR spectra has to be predicted or a fast and easy check of a structure is necessary, this approach is a very good opportunity. Moreover the large amount of disk space for saving the database or long time for loading data from external computers are no longer necessary. It would also be possible to perform the training of the network interactively, so that every scientist could create a network specialized in the groups of substances he or she is dealing with.

ACKNOWLEDGMENT

J.M. thanks the "Fond der chemischen Industrie" for a Kekulé stipend.

REFERENCES AND NOTES

(1) Sattler, M.; Schleucher, J.; Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93−158.
(2) Munk, E. M. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997−1009.
(3) Lindel, T., Junker, J., Köck, M. NMR-Guided Constitutional Analysis of Organic Compounds Employing the Computer Program COCON. *Eur. J. Org. Chem.* **1938**, *3*, 573−577. (A cocon version is available in the Internet under *http://cocon.org.chemie.uni-frankfurt.de*.)
(4) Peng, C.; Yuan, S. G.; Zheng, C. Z.; Hui, Y. Z.; Wu, H. M.; Ma, K.; Han, X. W Application of expert system CISOC-SES to the Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 814−819.
(5) Peng, C.; Bodenhausen, G.; Qiu, S. X.; Fong, H. H. S.; Farnsworth, N. R.; Yuan, S. G.; Zheng, C. Z. Computer-assisted structure elucidation: Application of CISOC-SES to the resonance assignment and structure generation of betulinic acid. *Magn. Reson. Chem.* **1998**, *36*, 267−278.
(6) Nuzillard, J. M. Computer-assisted structure determination of Organic Molecules. *J. Chim. Phys.-Chim. Biol.* **1998**, *95*, 169−177.
(7) Pihlaja, K.; Kleinpeter, E. *Carbon-13 NMR Chemical Shifts in Structural and Stereochemical Analysis*; VCH Verlagsgesellschaft: Weinheim, 1994.
(8) Bremser, W. HOSE − a novel substructure code. *Anal. Chim. Acta* **1973**, *103*, 355−365.
(9) *SpecInfo database*; Chemical Concepts, STN: Karlsruhe.
(10) Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A. *Carbon-13 NMR Spectral Data*; Verlag Chemie: Weinheim, 1981.
(11) Robien, W. *CSEARCH*; http://felix.orc.univie.ac.at/~wr/csearch_server_info.html.
(12) Trepalin, S. V.; Yarkov, A. V.; Dolmatova, L. M.; Zefirov, N. S.; Finch, S. A. E. WINDAT − an NMR Database Compilation Tool, User Interface, and Spectrum Libraries for Personal Computers. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 405−411.
(13) *CNMR database*; Advanced Chemistry Development Inc.: 133 Richmond Street West, Suite 605 Toronto, Ontario, Canada M5H 2L3.
(14) Schindler, M.; Kutzelnigg Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules. W. *J. Chem. Phys.* **1982**, *76*, 1919−1933.
(15) Pretsch, E.; Clerc, J. T.; Seibel, J.; Simon, W. *Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden*; Springer-Verlag: Berlin, 1981.
(16) Fürst, A.; Pretsch, E.; Robien, W. Comprehensive Parameter Set for the Prediction of the ^{13}C NMR Chemical Shifts of sp3-hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1990**, *233*, 213−222.
(17) Pretsch, E.; Fürst, A.; Robien, W. Parameter set for the Prediction of the ^{13}C NMR Chemical Shifts of sp^2- and sp-hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1991**, *248*, 415−428.

(18) Ewing, D. 13C Substituent Effects in Monosubstituted Benzenes. *Org. Magn. Reson.* **1979**, *12*, 499−524.
(19) Thomas, S.; Strohl, D.; Kleinpeter, E. Computer Application of an Incremental System for Calculating the 13C NMR Spectra of Aromatic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 725−729.
(20) Gloor, A., Cadisch, M.; Bürgin-Schaller, R.; Farkas, M.; Kocsis, T.; Clerc, J. T.; Pretsch, E.; Aeschimann, R.; Badertscher, M.; Brodmeier, T.; Fürst, A.; Hediger, H.-J.; Junghans, M.; Kubinyi, H.; Munk, M. E.; Schriber, H.; Wegmann, D. *SpecTool: A Hypermedia Book for Structure Elucidation of Organic Compounds using Spectroscopic Methods*; Chemical Concepts: Weinheim, 1994.
(21) Cheng, H. N.; Kasehagen, L. J. Integrated Approach for C-13 Nuclear Magnetic Resonance Shift Prediction, Spectral Simulation and Library Search. *Anal. Chim. Acta* **1994**, *285*, 223−235.
(22) Meiler, J.; Meusinger, R.; Will, M. Neural Network Prediction of ^{13}C NMR Chemical Shifts of Substituted Benzenes. *Monatsh. Chem./ Chem. Monthly* **1999**, *130*, 1089−1095.
(23) Burns J. A.; Whitesides, G. M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev.* **1993**, *93*, 2583−2601.
(24) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of Recurrent Neural Networks in Chemistry. Prediction and Classification of 13C NMR Chemical Shifts in a Series of Monosubstituted Benzenes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747.
(25) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of neural networks with feedback connections in chemistry: prediction of 13C NMR chemical shifts in a series of monosubstituted benzenes. *J. Mol. Struct. (Theochem.)* **1992**, *277*, 87−107.
(26) Sklenak, S.; Kvasnicka, V.; Pospichal, J. Prediction of 13C NMR Chemical Shifts by Neural Networks in a Series of Monosubstituted Benzenes. *Chem. Papers* **1994**, *48*, 135−140.
(27) Doucet, J. P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. Neural networks and 13C NMR shift prediction. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324.
(28) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157−1164.
(29) Svozil, D.; Pospichal, J.; Kvasnicka, V. Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924−928.
(30) Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to multilayer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 43−62.
(31) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D. 13C NMR chemical shift sum prediction for alkanes using neural networks. *Computers Chem.* **1997**, *21*, 437−443.
(32) Panaye, A.; Doucet, J. P.; Fan, B. T.; Feuilleaubois, E.; Azzouzi, S. R. E. Artificial neural network simulation of 13C NMR shifts for methyl-substituted cyclohexanes. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 129−135.
(33) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. 13C NMR Chemical Shift Prediction of sp2 Carbon Atoms in Acyclic Alkenes Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653.
(34) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. 13C NMR Chemical Shift Prediction of the sp3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 587−598.
(35) Thomas. S.; Kleinpeter, E. Assignment of the ^{13}C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network. *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504−507.
(36) Clouser, D. L.; Jurs, P. C. Simulation of the 13C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks. *Carbohydr. Res.* **1995**, *271*, 65−77.
(37) Clouser, D. L.; Jurs, P. C. The simulation of 13C nuclear magnetic resonance spectra of dibenzofurans using multiple linear regression analysis and neural networks. *Anal. Chim. Acta* **1996**, *321*, 127−135.
(38) Clouser, D. L.; Jurs, P. C. Simulation of the 13C Nuclear Magnetic Resonance Spectra of Ribonucleosides Using Multiple Linear Regression Analysis and Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168−172.
(39) Meusinger, R.; Moros, R. Determination of Quantitative Structure-Octane Rating Relationships of Hydrocarbons by Genetic Algorithms. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 67−78.
(40) Meiler, J. Smart; http://www.krypton.org.uni-frankfurt.de/~mj, 1998.
(41) Meiler, J. C_Shift; http://www.krypton.org.uni-frankfurt.de/~mj, 1999.
(42) Höfle, G.; Bedorf, N.; Steinmetz, H.; Reichenback, H.; Gerth, K. Epothilon A and B - Novel 16-membered macrolides with cytotoxic activity: Isolation, crystal structure, and conformation in solution. *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1567−1569.

CI000021C