communications in mathematical
match
and in computer chemistry

no. 45, March 2002

MATCDY (45) 85 - 108 (2002)

ISSN 0340 - 6253

Ranking MOLGEN Structure Proposals by ^{13}C NMR Chemical Shift Prediction with ANALYZE

JENS MEILER*,[1], MARKUS MERINGER[2]

[1]HHMI at the University of Washington, Box 357350, Seattle, WA 98195-7350, USA
[2]Dep. of Mathematics, University of Bayreuth, D-95440 Bayreuth, Germany

Key words. structure generator; neural networks; ^{13}C chemical shift; automated structure elucidation; NMR; MOLGEN; ANALYZE;

*Corresponding author

Contact address:

University of Washington
Box 357350
Seattle
WA 98195-7350
USA

Tel.: +1 (206) 543 7134
Fax.: +1 (206) 685 1792
Mail: jens@jens-meiler.de

Abstract

Artificial neural networks are capable of predicting the ^{13}C chemical shifts of organic molecules nearly as fast as incremental methods while maintaining the accuracy of database methods. In this article, we apply a recently developed neural network (Meiler et. al., *J. Chem. Inf. Comput. Sci.* 2000, *40*, 1169-1176), to the screening of large sets of molecules obtained by structure generators in the process of automated structure elucidation. Specifically, we apply the network to sets of structures generated by MOLGEN (Benecke et. al., *Anal. Chim. Acta* 1995, *314*, 141-147) for ten randomly selected molecules of less than 13 non-hydrogen atoms. The computed ^{13}C NMR spectra are compared to the experimental spectrum; in all cases, the computed spectrum belonging to the example molecule yields a significantly smaller deviation to the experimental data then all other predicted spectra. This result suggests that the approach is suitable for automated structure prediction for organic molecules with up to 12 non-hydrogen atoms.

Introduction

Structure elucidation is one of the basic needs in organic chemistry after a substance is synthesized or isolated. A large variety of powerful methods such as multidimensional high resolution NMR spectroscopy or mass spectroscopy is available for this purpose. Databases contain spectra of hundreds of thousands of organic compounds and allow a fast comparison of a new spectrum with all existing spectra in order to find similarities or identities. However, the number of substances to be analyzed is also increasing rapidly, which creates a need for partially or fully automated approaches to structure elucidation.

During the search for pharmaceutical drugs or other biological agents more and more substances are synthesized. Modern techniques, combinatorial methods and fully automated synthesis further increase the number of samples. Also the measurement of analytical data can be done in a highly automated manner, so that the manpower-intensive structure elucidation becomes the bottleneck of the "structure elucidation pipeline".

Using NMR spectroscopy, one way of finding the constitution of a compound is to suggest a structure and to test whether this suggestion is compatible with all the data derived from the NMR spectra, including chemical shift, multiplicity and connectivity information from Heteronuclear Single-Quantum Coherence (HSQC) spectra. Usually this process has to be repeated until a suggested constitution is compatible with every experiment and ideally all other possibilities should have been excluded. However, the exclusion of all other theoretically possible structures is a challenging task since it includes the discussion of every single possible constitution for a given molecular formula. The number of possible constitutions becomes huge already for substances with about twelve non-hydrogen atoms. Although the chemist can use his knowledge to exclude large parts of the structural space in a first step (for example all substances containing a carbonyl if there is no ^{13}C chemical shift higher than 150 ppm), for more complex cases it quickly becomes impossible to survey the space of all possible constitutions.

At this point a structure generator is needed to generate all structures that fulfill a certain set of boundary conditions (e.g. molecular formula and optional information on the H-distribution, hybridization and substructures (MOLGEN)(1) or molecular formula and connectivity information from NMR spectra (COCON)(2)). The advantage of having all possible structures at hand comes often along with the disadvantage of a large set of data that cannot be analyzed manually.

^{13}C NMR chemical shift data are especially sensitive to the constitution of an organic compound, since the chemical environment of every single carbon atom in the molecule is described by such a number. Since carbon is the most common non-hydrogen atom in organic compounds and is involved in intermolecular interactions only to a limited amount, the ^{13}C NMR chemical shift represents almost pure, noise-free connectivity information. If it is possible to predict the carbon chemical shift from the constitution of a molecule quickly and accurately, an automated ranking of the structure generator results becomes possible. Consequently the prediction of ^{13}C chemical shifts plays an important role in structure elucidation. Two basically opposite approaches are *ab initio* and *empirical* calculations.

Ab initio calculations compute magnetic properties from first principles, as the mixed second derivative of the energy with respect to an applied magnetic field and the nuclear magnetic moment [e.g., Schindler and Kutzelnigg(3)]. Starting from a three-dimensional structure of the compound under consideration highly accurate results can be produced for the entire molecular system. However, the necessity to predetermine both the constitution and the correct configuration/conformation restricts the applicability of this calculation method. The correct three-dimensional structure is often unknown and multiple conformations have to be taken into account for small and flexible molecules, particularly. Extensive optimization of the spatial structure on a high level and/or consideration of multiple conformations render such calculations very time-consuming and expensive. On the other hand, the resulting chemical shift values are not affected by previous experimental results and are thus more impartial. Especially for strained and other unusual systems, chemical shift values are often predicted more accurately by using *ab initio* calculations.

By contrast *empirical* approaches rely on knowledge of chemical shifts for a large set of known molecular structures. The first publications introducing the approach known as “increment method” were published by Grant and Paul already in 1964(4), by Lindeman and Adams in 1971(5) and by Clerc and Sommerauer in 1977(6). The advantage of the method is its simplicity that allows the transfer to nearly every class of substances and a straightforward calculation of the shift values even by hand. These methods are still under development (7) and can be applied to all ordinary organic substances. However, the limitation of this simple approach is that all interactions between several substituents of a carbon atom are ignored. Therefore large deviations between experimental and predicted chemical shift values are often obtained for highly substituted fragments.

Soon after computers became available to the general public, ^{13}C NMR spectra were stored in databases (e. g. SPECINFO(8) or CSEARCH(9)) to serve for extensive data analysis. Bremser et. al.(10) introduced a *h*ierarchically *o*rdered *s*pherical description of *e*nvironment (HOSE) code to describe the constitutional environment of a carbon atom. The longer the code the more spheres are described. Lists of such descriptions covering the first up to 5 spheres around a carbon atom were stored together with the corresponding chemical shift information. Now for every molecule prediction of ^{13}C NMR chemical shift is possible by calculating the HOSE code for each carbon atom and a subsequent search through the database for similar codes. This method is known to provide a very exact prediction of the carbon chemical shift if the database contains similar HOSE codes. One obvious advantage of

HOSE code prediction is the reference to all original data enabling a direct check of the assignment. Disadvantages of the method are a relatively slow prediction compared to increment methods, the necessity of access to the large database and an enhanced uncertainty for structures outside the space covered by the database.

With the introduction of artificial neural networks to chemistry (11) in recent years, their potential for ^{13}C NMR chemical shift prediction was evaluated. At first, similar to increments they were applied to restricted classes of substances (12-18); later, approaches were introduced that cover the entire space of organic compounds(9,19,20). Artificial neural networks combine the advantages of increments and HOSE code prediction: They are fast (once the networks are trained), precise (since interactions between substituents are considered), independent from direct access to a database, and (compared to HOSE code and increment methods) especially accurate in estimating chemical shifts of newly synthesized molecules that are badly represented in the database.

We discuss in this paper an application of our previously introduced neural network ^{13}C NMR chemical shift prediction (19) as an efficient filter for a structure generator. The program ANALYZE(21) provides the comparison of a given experimental NMR spectrum with neural network predicted NMR spectra for a set of given structures and ranks the structures with respect to the similarity between experimental and computed data.

Recently we combined ANALYZE with COCON showing that it is possible to extract a small amount (~0.1%) of probable constitutions out of a complete set of possible constitutions for proton-poor compounds with up to 25 non-hydrogen atoms(21). Moreover, it was possible to use the quality measure of the similarity between an experimental and a computed ^{13}C NMR spectrum for a suggested constitution as fitness function of a genetic algorithm (GENIUS (22)). This genetic algorithm is taking the role of a structure generator by creating populations of constitutions that evolve under the selection pressure of the fitness function. Therefore the constitution is optimized to fulfil the experimental ^{13}C NMR spectrum. This algorithm was proven to solve the constitution of molecules with up to 20 non-hydrogen atoms automatically.

However, COCON relies on connectivity information, which implies the record of more experimental data in time-consuming higher dimensional NMR experiments. GENIUS generates only a part of the complete constitutional space and may therefore miss the correct solution. For small organic molecules of 12 non-hydrogen atoms, the calculation of all possible constitutions is at the limit of computational power today.

MOLGEN is a powerful structure generator that computes, starting from a molecular formula and optional further conditions all possible constitutions rapidly and free of redundancy (1,23,24). By applying the ^{13}C NMR chemical shift filter on complete sets of MOLGEN structures, we want to address three questions in this paper:

- How efficient is the ^{13}C NMR chemical shift comparison applied to sets of structures that cover the structure space of one molecular formula completely?
- Up to which size of molecules does this combination yield a practical and reliable method for automated structure elucidation?
- How can one early recognize such parts of the structure space that need not to be generated since they do not contain the correct solution?

The latter point is of special interest for applying this method to compounds of a more realistic size at the scale of today's organic synthesis.

Methods

(MOLGEN:) This generator of structural formulae knows two generation methods: *orderly* and *restricted* generation. While restricted generation is able to process various structural restrictions efficiently, orderly generation is recommended, if only the molecular formula is given, or in addition several restrictions such as hybridizations or a hydrogen distribution. As already mentioned, we have input only the molecular formula and therefore orderly generation was used.

MOLGEN calculates chemical constitutions as connectivity matrices. Filling the n x n connectivity matrix in all possible ways according to the given molecular formula with n atoms is no serious algorithmic problem, but is very time-consuming for increasing n. The second problem is to avoid redundancy in the output, the so-called *isomorphism problem*; i.e., we must decide which connectivity matrices represent identical constitutions. Naively, this problem has time complexity $O(n!)$, because in the worst case one would have to apply all the $n!$ permutations of the symmetric group in order to decide whether two connectivity matrices are isomorphic. With the aid of combinatorics, algebra and group theory, these problems are cut down immensely.

Mathematically we identify constitutions with *unlabeled molecular graphs*. A molecular graph is an undirected multigraph together with a coloring of the vertices, which represents the atoms' chemical elements. Unlabeled molecular graphs on n vertices are the orbits of the group action of the symmetric group S_n on the labeled molecular graphs on n vertices. In mathematical terms the problem is to find a full (but non-redundant) set of orbit representatives of this group action. A very efficient method to solve this problem is Read's *orderly generation* (25), in which structures are enlarged successively by adding edges. A linear order is introduced on the objects and the minimal structures in each orbit are defined to be the canonical orbit representatives. One can prove that minimal orbit representatives arise from stepwise enlargement of already minimal predecessors. Therefore, whenever a non-minimal graph is reached, we do not need to insert further edges, because this will not lead to a minimal orbit representative. This technique already reduces the computational effort enormously. Further details about the algorithm can be found in the theses of Grund (26) and Grüner (27).

(ANALYZE:) The neural network approach for predicting ^{13}C chemical shifts is described in detail elsewhere (19) and therefore will be only summarized briefly here. From the SPECINFO database ~100 000 organic molecules (containing exclusively H, C, N, O, S, P, halogens) with known ^{13}C NMR spectrum were selected. Out of this set of molecules a training set (95%), a monitoring set (2%) and an independent set (3%) of molecules were randomly picked.

The constitutional environment of every single carbon atom was described using up to 1,696 numerical descriptors: These parameters encode every single substituent of the carbon atom of interest within the first three spheres (13 atoms x 8 properties = 104 parameters). For all further spheres, only the number of atoms that belong to a special atom type (32 atom types were previously defined using element number, period, hybridization and number of bond hydrogen atoms (19)) is determined whereas all atoms that belong to sphere eight an higher are combined in one sum sphere. This procedure results in 160 (= 32 atom types x 5 spheres) additional input parameters which are incorporated twice, once counting all atoms and a second time only considering atoms that belong to a conjugated π electronical system with the carbon atom of interest. This leads to 424 (= 104 + 160 + 160) parameters for a single subtituent and therefore to 1,696 (= 4 x 424) parameters for a quaternary carbon atom. Nine different neural networks were trained to predict the chemical shift for the nine defined carbon atom types (〉C〈, 〉CH–, –CH_2–, –CH_3, =C〈, =CH– / =CH_2, ≡C– / ≡CH / =C=, 》C–, 》CH). Standard three layer feed forward neural networks containing up to 1,696 input units, 32 hidden neurons and one output (up to 54,337 weights) were trained using the back-propagation algorithm with a total of ~1,300,000 carbon atom environments out of the training set of data until the RMSD of the monitoring set of data was minimized. For the independent set of data, a standard deviation of 2.4 ppm and a mean deviation of 1.6 ppm was obtained.

For the present investigation 10 molecules were randomly selected out of the independent set of data that contain 9, 10, 11, and 12 non-hydrogen atoms. Using the program MOLGEN (1,23) all possible constitutions for these molecules were generated using only the molecular formula as input. For every compound in each of the resulting sets of data the ^{13}C NMR chemical shift spectrum was computed using the neural network approach and compared with the experimental data. The RMSD (root mean square deviation) of the computed and the experimental chemical shift values was calculated after the list of carbon atoms was sorted with respect to an increasing shift value for the experimental data as well as for the computed ones. Finally all structures were ranked starting with the lowest RMSD value.

Table 1 shows the selected compounds, gives some detail of the MOLGEN and ANALYZE runs and presents their results. The structure generation with MOLGEN 3.5 was done on a 800MHz Pentium III PC under Windows NT. Using orderly generation, the size of main memory has no influence on the performance. The ANALYZE calculations were performed on 12 PC equipped with two 1GHz Pentium III processors with 1GB main memory running in a cluster under Linux. The computation time for ANALYZE given in Table 1 includes reading and writing as well as the additional data handling. The time necessary for the computation of the chemical shift alone – once the molecule is read – is about five-fold faster.

Table 1: Molecular and constitutional formula, computational aspects, and results obtained for the ten example compounds

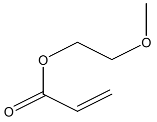
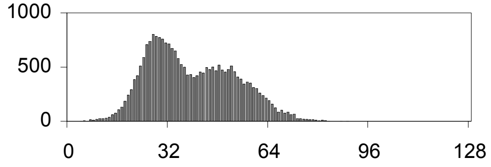
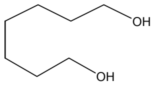
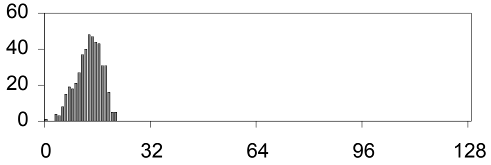
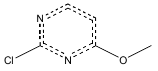
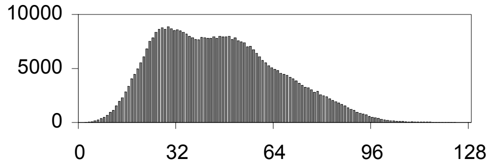
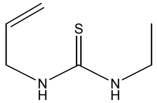
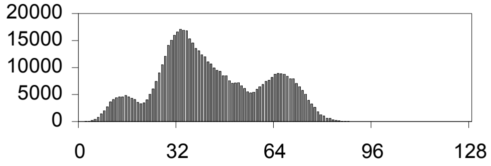
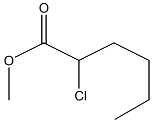
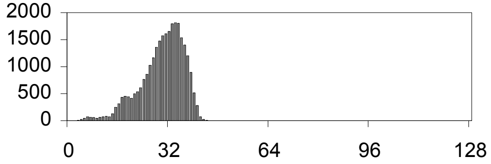
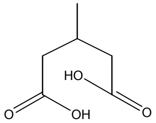
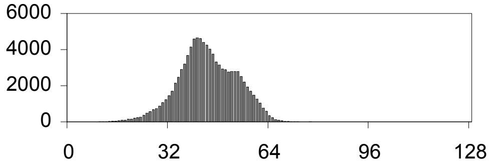
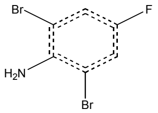
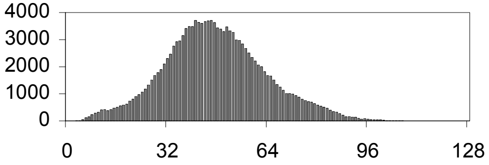
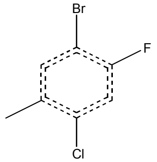
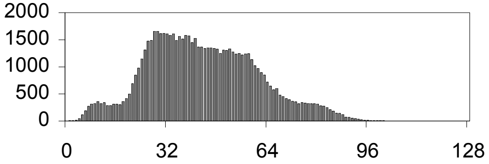
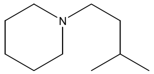
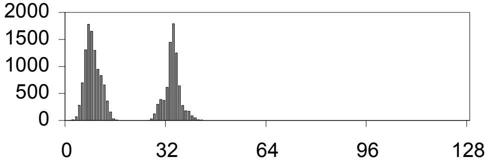
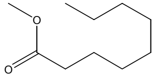
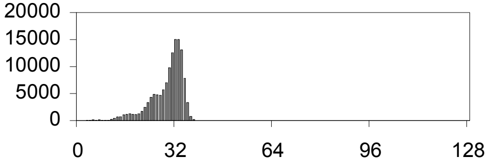
Compound				Computation			Shift comparison			
Nr	Name	Molecular formula	Constitutional formula	Number of generated structures	Time MOLGEN (s)	Time ANALYZE (s)	Best/ correct RMSD (ppm)	2^{nd} best RMSD (ppm)	Worst RMSD (ppm)	Distribution of ^{13}C NMR chemical shift RMSD values (ppm)
1	Methoxyethyl acrylate	$C_6H_{10}O_3$		23 838	>1	412	0.52	1.94	89.51	
2	Heptane-1,7-diol	$C_7H_{16}O_2$		463	>>1	6	1.00	3.39	21.17	
3	2-Chloro-6-methoxypyrimidine	$C_5H_5N_2OCl$		447 891	22	6 291	0.97	1.30	129.3	

Table 1 (continued): Molecular and constitutional formula, computational aspects, and results obtained for the ten example compounds

Compound				Computation			Shift comparison			
Nr	Name	Molecular formula	Constitutional formula	Number of generated structures	Time MOLGEN (s)	Time ANALYZE (s)	Best/ correct RMSD (ppm)	2^{nd} best RMSD (ppm)	Worst RMSD (ppm)	Distribution of ^{13}C NMR chemical shift RMSD values (ppm)
4	*N*-Allyl-*N'*-ethylthiourea	$C_6H_{12}N_2S$ $C_6H_{14}N_2S$ $C_6H_{16}N_2S$		709 259	41	8 213	0.70	1.65	93.50	
5	Methyl 2-chlorohexanoate	$C_7H_{13}O_2Cl$		27 575	>1	589	0.75	3.56	46.11	
6	3-Methylglutaric acid	$C_6H_{10}O_4$		97 394	3	1 447	1.42	4.04	82.33	

7	2,6-Dibromo-4-fluoroaniline $C_6H_4NFBr_2$		140 748	7	2 455	0.96	2.29	111.00	
8	5-Bromo-2-chloro-4-fluorotoluene C_7H_5FClBr		71 394	2	1 413	0.46	1.40	102.44	
9	*N*-isopentyl-piperidine $C_{10}H_{21}N$		17 884	>1	349	0.47	1.09	43.30	
10	Methyl nonanoate $C_{10}H_{20}O_2$		126 750	5	3 139	0.77	1.39	40.60	

Results and Discussion

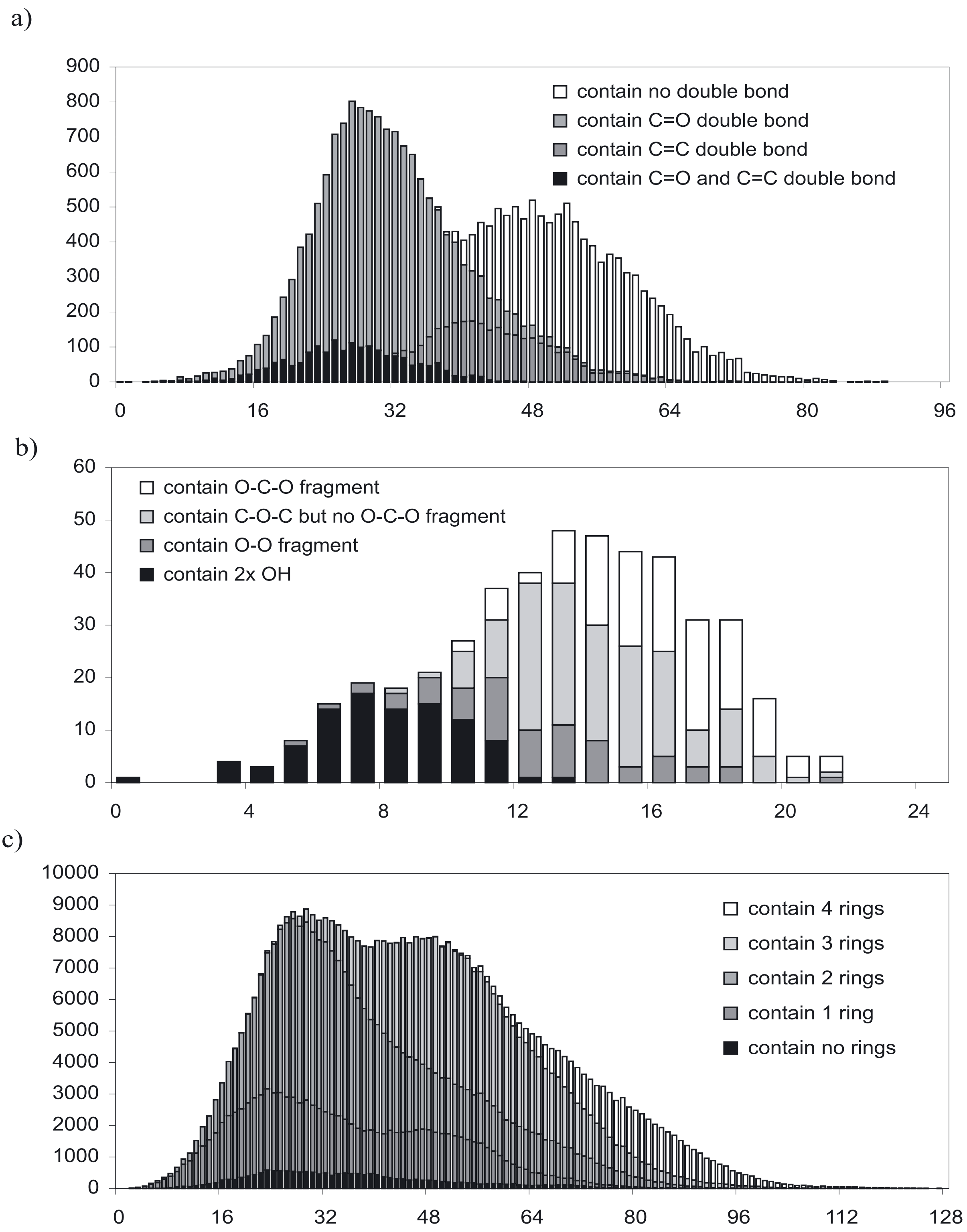
In each of the ten examples the presented method is able to rank the correct structure as first. Also, the difference between the first ranked and all other structural proposals is significant. The difference between the first and the second ranked structures lies between 0.33 ppm and 2.81 ppm and tends to become smaller with increasing size of the set of molecules, as one would expect. Furthermore, we know from the application of the chemical shift prediction on larger molecules (CoCon (2,21) and Genius (22)), that with an increase in the number of possible constitutions, solutions with a smaller deviation to the experiment than the true solution (false positives) will also occur. The occurrence of such false positives depends on the computational and the experimental error as well as on the size of the set of molecules. The number of non-hydrogen atoms is not necessarily a good measure for the size of the structural space. In the application of Genius 14 non-hydrogen atoms were necessary to obtain a false positive behavior for the first time.

The distributions of RMSD values have a very different shape for the ten examples, often showing more than just one maximum. This behavior suggests that the set of compounds can be subdivided into several subsets. Presumably the subset that contains the correct structure produces also the lowest average deviation. All other subsets differ in one structural feature, which is likely to change the NMR spectrum for all members of this subset in about the same manner, and makes it impossible to agree with the experiment. However, the number of such subsets varies and also the separation of the subsets changes dramatically when looking at different examples. The fourth example shows three well-defined maxima but the distribution around these maxima overlap, whereas example nine shows two completely separated distributions; by contrast, in example seven, it is hard to recognize more than one maximum, although the distribution seems to include some shoulders. However, the occurrence of such well defined subspaces would be a starting point for a gentle reduction of the structural space to be generated, if the responsible structural features can be detected prior.

1. The first set of structures contains 23,838 members. Due to the three hetero atoms and the two double bond equivalents the number of possible constitutions is in the medium range for this molecular formula. The distribution of the RMSD to the target spectrum shows two maxima at around ~30 and ~50 ppm. While the first subset contains structures with one or two double bonds, the second subset contains structures without double bonds, i.e. bicyclic structures. The few candidates with deviations smaller than 6 ppm are the structures containing both, a C=C and a C=O bond. Figure 1a) illustrates this analysis by coloring the distribution plot according to the occurrence of double bonds. It is easy to see that structure proposals without double bonds and with only one C=C bond do not achieve low RMSD values at all. One C=O double bond is absolutely necessary to achieve a RMSD smaller than 30 ppm. An RMSD value of 6 ppm is the lower limit for structures that contain only one C=O bond and no additional C=C bond. Therefore it is very likely, that the correct constitution contains these both structural features, which might also be guessed directly from looking at the experimental ^{13}C NMR spectrum.

2. The second set of structures has 463 members, and is the smallest of the ten generated sets. The influence of the number of double bond equivalents on the number of possible constitutions is impressive. In comparison with the previous example, only one oxygen was replaced by a carbon atom and the two double bond equivalents were deleted. We will use this relatively small set to look at the structure proposals with respect to their RMSD value to the experimental spectrum in some more detail. Figure 1b) visualizes the distribution of four subsets containing all substances with a O–C–O fragment (a), with a C–O–C fragment (b), with a O–O fragment (c) and finally all structures that contain two OH groups (d). Figure 2 shows the members of each subset with the lowest and the highest RMSD value

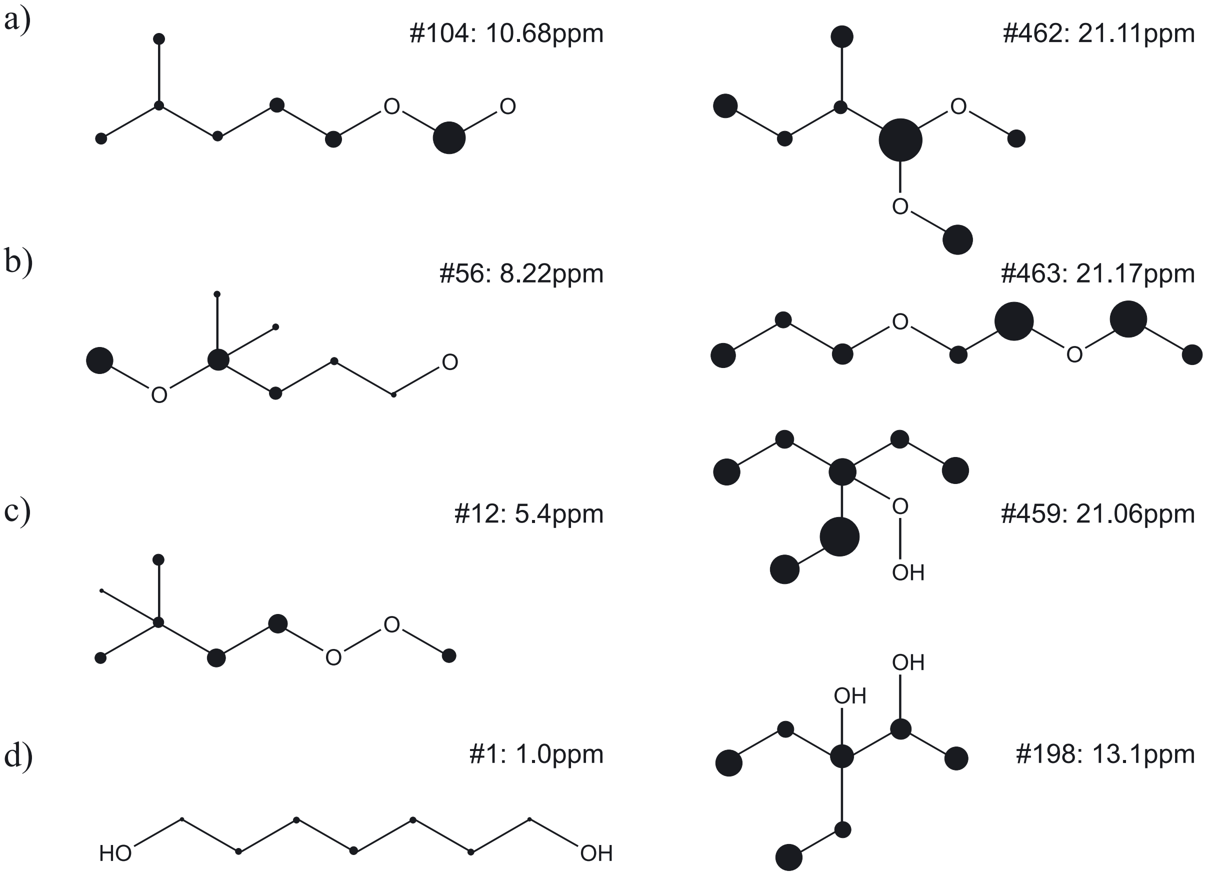
Figure 1: Distribution of the ^{13}C chemical shift RMSD values computed from the experimental and the neural network computed spectrum. On the x axis the RMSD value in ppm is given and on the y axis the number of structures with this deviation are counted. Diagrams a), b) and c) correspond to the examples 1, 2, and 3 in table 1 and in the text.



to the target spectrum. As discussed for the first example, it is again seen that the correct solution is clearly preferred by a low RMS value and the whole subset of structures containing the same structural features has a low RMSD compared to the other subsets. The subset containing all structures with a O–O fragment comes closest due to the comparable chemical shifts of the attached carbon atoms (see Figure 2c left). If there is only one carbon attached to the fragment, the RMSD becomes huge (Figure 2c right). All members of the other two subsets contain at least one carbon atom with a chemical shift value that is far too high. In case b) – structures that contain C–O–C – at least three carbon atoms are attached to an oxygen and in case a) – structures that contain O–C–O – the chemical shift of the carbon between the two oxygen atoms does not agree with the obtained experimental data.

3. The third example contains four double bond equivalents due to its aromatic structure. The large number of hetero atoms increases the number of possible constitutions further. Almost 450,000 different structural formulae are compatible with this molecular formula. Nevertheless, the structure that belongs to the experimental NMR spectrum is still correctly identified. Only a small fraction of these structures (365) is aromatic, yielding RMSD values between 1 ppm and 38 ppm with a maximum in the distribution at 20 ppm. It is easy to imagine that many polycyclic structures become theoretically possible in such a situation, although they are unlikely to exist in reality. To investigate this behavior, we analyze the distribution of molecules with respect to the number of rings within every structure (Figure 1c). As expected, structures with more than two rings do not achieve low RMSD values at all, since they do not contain two C=C double bonds necessary to exhibit the characteristic four shift values in the olefinic/aromatic region. Structures with zero, one or two rings can form two C=C double bonds but need not. Therefore some members out of these subsets yield low RMSD values but the distribution is very broad. While the member with the lowest RMSD of the one ring subset is the correct structure, the lowest RMSD structure

Figure 2: Structures with lowest and highest RMSD to the target spectrum for the four subspaces generated for Example 2 (compare Figure 1b and text), respectively. The area of the black circles at every carbon atom position are proportional to the chemical shift deviation. The given identification number corresponds to the ranking in the chemical shift comparison.



without rings achieves 7.5 ppm and the lowest RMSD structure with two rings achieves 3.6 ppm. Also the number of carbon atoms that have a sp^2 hybridization as indirectly applied in example 1 would be an efficient filter for this set of molecules.

With these three examples the possibilities to analyze sets of molecules of different sizes in context with a given ^{13}C NMR spectrum are summarized. Due to the nature of a ^{13}C NMR spectrum – giving every carbon atom one chemical shift value – filters that look for certain fragments, the number of rings and also only the number of carbon atoms in a certain hybridization state are suitable criteria to investigate the generated structure space. Such filters might also be applied during the generation of structural spaces too large to be generated completely. Since we do not want to discuss a certain problem in detail but introduce a general method for handling such sets of data, we describe the remaining seven examples only briefly.

The next set of structures is the largest out of the ten presented here. Again the combination of a few hetero atoms with two double bond equivalents yields no less than 709,259 possible structures. It was necessary to exclude ~20% of these structures since they contain sulfur with 3 to 6 single bonds. These atom types are not defined for the ^{13}C chemical shift prediction since their occurrence in the SPECINFO database are so rare that the training of the corresponding neural network connections was impossible (19). As one would intuitively guess, the first clearly separated subset contains structures with three sp^2 – carbons, the second largest subset can be split into structures with two or four sp^2 – carbons, the comparably small set of structures with one sp^2 – carbon comes next at ~50 ppm maximum and the last subset containing molecules with no sp^2 – carbon atoms has its maximum at ~70 ppm.

The next set of molecules is comparably small (27,575 molecules) and most of its members are ranked with high RMSD values. The unique chemical shift of the carbonyl is only achieved by a few substances that really contain a carbonyl. The next local maximum in the distribution corresponds to structures that contain a C=C double bond preferably with at least one oxygen as direct substituent, and the biggest peak is the center of the subset of all structures not containing any sp^2 – carbon atom.

An even more efficient suppression of wrong answers is obtained for the next example. Only ~1% of all generated structures contain two carbonyls or have a –O–C=C–O– fragment and achieve therefore low RMSD values.

The next two example sets are both very large and contain mostly non-aromatic structures. Again the number of sp^2 – carbon atoms is a good filter. Since nitrogen can participate in a double bond, in example 7 structures with six, five, four, three, two, one and no sp^2 – carbons are possible yielding multiple heavily overlapping subsets. The average deviation increases with decreasing number of sp^2 – carbons.

Example 8 does not allow all integer numbers of sp^2 – carbons between six and zero, but only the even numbers six, four, two and zero because no alternative partner for double bonds is available in contrast to the previous case. The overall set of possible structures becomes smaller, and only four subsets (instead of seven in example 7) become better separated as indicated by the four maxima in the distribution plot. This is also the example where on can observe one general rule most impressively: The four maxima caused by the number of sp^2 – carbons are about equidistant. Starting with ~12 ppm one can add ~20 ppm to yield the position of the next maximum. This makes sense, since the deletion of one double bond yields a decrease of the two chemical shift values for the two carbon atoms, that should be constant in average.

The molecular formula of example 9 contains only one nitrogen and one double bond equivalent. These facts limit the number of possible constitutions to be 17,884. The distribution shows two well separated subsets of molecules at an average RMSD of about 10 ppm and 35 ppm. While the first subset contains all molecules with no double bonds, the introduction of any double bond (C=C or C=N) yields a RMSD larger than 26 ppm. Within these two subsets further differences can be obtained. The first subset contains tertiary, secondary and primary amines in this order while the second group contains two major subsets with either a C=C or a C=N fragment.

The last example deals again with a larger set of data (126,750 structures), not due to a high number of double bond equivalents, but due to the higher number of now 12 non-hydrogen atoms. The distribution is similar to those in examples 5 and 6 where carbonyl atoms were present. Only a small fraction of all proposals contain this fragment or a C=C–O fragment and are therefore able to achieve a low RMSD value. All other molecules end up with large RMSD values in the distribution plot.

Although in all ten examples the correct structure for the experimental spectrum was picked, we know that this is not the case if the structure size exceeds certain limits. The question is: How probable is it to find a structural proposal with a lower deviation to the target spectrum than the true structure itself? Since the average deviation for the chemical shift prediction is 1.6 ppm, such a structure must be usually below this limit. The probability that this happens depends not only on the number of possible structures but also on the position of the unknown structure in the structural space with respect to the ^{13}C NMR spectrum. If the region is very dense and a lot of structures with similar spectra exist (e.g. examples 2, and 9) the probability rises while in regions with only few structures (e.g. examples 1, 5, 6, and 10) the probability is lower. The size of subspaces that contain all substances with similar spectra is a measure for this probability.

Although the method is not practicable at present for most molecules with more than 12 non-hydrogen atoms for the reason of high computation times, intelligently chosen boundary conditions for the structure space to be generated would circumvent this problem. Instead of applying the filter afterwards as done in the discussed approach, the structure space needs to be decreased before the start of the calculation (e.g. by defining a certain number of carbonyl or more generally sp^2 – carbon atoms) or ideally and more specifically on the fly. During the MOLGEN computation the spectrum of the generated structures is computed, compared with the target and the result is used to decide which regions of the structure space are generated (first).

Both approaches are already used. COCON generates only a predefined part of the overall structure space. However, it relies on experimentally expensive two dimensional connectivity information. GENIUS determines the structure space to be generated on the fly and has proven to be very efficient in doing this. However, since it is a genetic algorithm there will be never a guarantee that it really generates all members of a subspace.

A combination of GENIUS and MOLGEN might be very efficient. GENIUS finds structures quickly that are similar to the correct one (searching the structure space) but it converges very slowly in the end of the computation (local minimization). Using the preliminary rapidly accessible GENIUS result as starting point a MOLGEN run could evaluate the size of the local minimum and compute all members to achieve the final local minimization.

Conclusion

It was demonstrated that the combination of computing the complete structural space covered by one molecular formula with a subsequent neural network based prediction of ^{13}C chemical shift values allows the unambiguous determination of the correct structure to a given ^{13}C NMR spectrum for several example compounds with up to twelve non-hydrogen atoms. For all ten example molecules, the method yields a substantially lower RMSD of the predicted versus experimental chemical shift values for the correct structure compared with all other structures in the structural space. Considering the large size of the structural spaces with up to 700,000 structures, this result proves again that a ^{13}C NMR spectrum is a unique fingerprint for organic compounds of this size. On the basis of the experimental ^{13}C NMR spectrum a definition of subspaces becomes possible that have a high or a low probability to contain the structure to the corresponding NMR spectrum. To obtain these results, the structure generator MOLGEN was combined with the subsequent chemical shift prediction using the program ANALYZE. The overall calculation time for the examples was between 6 s and 8254 s. A further improvement of the method with the aim of targeting larger molecules by lower computation times should be achievable by an earlier incorporation of the experimental data to decrease the structure space being generated.

Acknowledgement

We would like to thank Bill Wedemeyer and C. Rücker for carefully reading the manuscript and useful discussions. J. M. would like to thank Martin Will and Reinhard Meusinger for their contribution to the earlier published chemical shift prediction method. M. M. thanks for financial support by the BMBF, contract 03C0318C.

References

(1) C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland, MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation, *Anal. Chim. Acta* **314**, 141-147 (1995).

(2) T. Lindel, J. Junker, and M. Köck, COCON: From NMR Correlation Data to Molecular Constitution, *J. Mol. Model.* **3**, 364-368 (1997).

(3) M. Schindler, and W. Kutzelnigg, Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules, *J. Chem. Phys.* **76**, 1919-1933 (1982).

(4) D. M. Grant, and E. G. Paul, Carbon-13 Magnetic Resonance. II. Chemical Shift Datat for the Alkanes, *J. Am. Chem. Soc.* **86**, 2984-2990 (1964).

(5) L. P. Lindeman, and J. Q. Adams, Carbon-13 Nuclear Magnetic Resonance Spectroscopy, *Anal. Chem.* **43**, 1245-1252 (1971).

(6) J.-T. Clerc, and H. Sommerauer, A Minicomputer Program Based On Additivity Rules For The Estimation Of 13C NMR Chemical Shifts, *Anal. Chim. Acta* **95**, 33-40 (1977).

(7) A. Fürst, and E. Pretsch, A computer program for the prediction of 13C NMR chemical shifts of organic compounds, *Anal. Chim. Acta* **229**, 17-25 (1990).

(8) "SpecInfo database", Chemical Concepts: Weinheim, 2001.

(9) W. Robien, Das CSEARCH-NMR-Datenbanksystem, *Nachr. Chem. Tech. Lab.* **46**, 74-77 (1998).

(10) W. Bremser, HOSE - A Novel Substructure Code, *Anal. Chim. Acta* **103**, 355-365 (1978).

(11) J. Zupan, and J. Gasteiger, "Neural Networks for Chemists", VCH Verlagsgesellschaft mbH: Weinheim, 1993.

(12) V. Kvasnicka, S. Sklenak, and J. Pospichal, Application of Recurrent Neural Network in Chemistry. Prediction and Classification of 13C NMR Chemical Shifts in a Series of Monosubstituted Benzenes, *J. Chem. Inf. Comput. Sci.* **32**, 742-747 (1992).

(13) O. Ivanciuc, Artificial neural networks applications. Part 6. Use of non-bonded van der Waals and electrostatic intermolecular energies in the estimation of 13C- NMR chemical shifts in saturated hydrocarbons, *Rev. Roum. Chim.* **40**, 1093-1101 (1995).

(14) D. Svozil, J. Pospichal, and V. Kvasnicka, Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes, *J. Chem. Inf. Comput. Sci.* **35**, 924-928 (1995).

(15) S. Thomas, and E. Kleinpeter, Assignment of the 13C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network, *J. Prakt. Chem./Chem.-Ztg.* **337**, 504-507 (1995).

(16) O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye, and J.-P. Doucet, 13C NMR Chemical Shift Prediction of sp2 Carbon Atoms in Acyclic Alkenes Using Neural Networks, *J. Chem. Inf. Comput. Sci.* **36**, 644-653 (1996).

(17) Z. Li, Y. Huang, F. Hu, Q. Sheng, and S. Peng, Neural Networks in spectroscopy. Estimation and prediction of chemical shifts of 13C NMR in alkanes by using subgraphs, *Bopuxue Zazhi* **14**, 507-514 (1997).

(18) J. Meiler, R. Meusinger, and M. Will, Neural Network Prediction of 13C NMR Chemical Shifts of Substituted Benzenes, *Monatshefte für Chemie* **130**, 1089-1095 (1999).

(19) J. Meiler, M. Will, and R. Meusinger, Fast Determination of 13C-NMR Chemical Shifts Using Artificial Neural Networks, *J. Chem. Inf. Comput. Sci.* **40**, 1169-1176 (2000).

(20) C. Le Bret, A General 13C NMR Spectrum Predictor using Data Mining Techniques, *SAR and QSAR in Environmental Research* **11**, 211-234 (2000).

(21) J. Meiler, E. Sanli, J. Junker, R. Meusinger, T. Lindel, M. Will, W. Maier, and M. Köck, Validation of Structural Proposals by Substructure Analysis and 13C NMR Chemical Shift Prediction, *J. Chem. Inf. Comput. Sci.* **ASAP on the WWW**, (2002).

(22) J. Meiler, and M. Will, Automated Structure Elucidation of Organic Molecules from 13C NMR Spectra using Genetic Algorithms and Neural Networks, *J. Chem. Inf. Comput. Sci.* **41**, 1535-1546 (2001).

(23) T. Wieland, A. Kerber, and R. Laue, Principles of the Generation of Constitutional and Configurational Isomers, *J. Chem. Inf. Comput. Sci.* **36**, 413-419 (1996).

(24) T. Grüner, A. Kerber, R. Laue, and M. Meringer, MOLGEN 4.0, *MATCH* **37**, 205-208 (1998).

(25) R. C. Read, Everyone a winner, *Annals of Discrete Mathematics* **2**, 107-120 (1978).

(26) R. Grund, Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten, *Bayreuther Mathematische Schriften* **49**, 1-113 (1995).

(27) T. Grüner, Strategien zur Konstruktion diskreter Strukturen und ihre Anwendung auf molekulare Graphen, *MATCH* **39**, 39-126 (1999).