Genius: A Genetic Algorithm for Automated Structure Elucidation from ^{13}C NMR Spectra
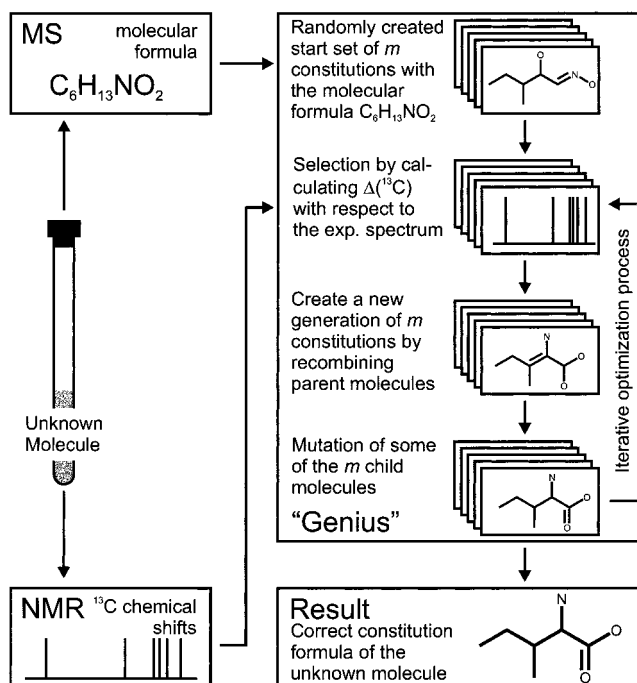
Jens Meiler*,† and Martin Will‡

University of Washington, Box 357350, Seattle, Washington 98195-7350, and BASF AG, Ludwigshafen, Germany

Received April 12, 2001; Revised Manuscript Received December 12, 2001

The routine work of structure elucidation of molecules produced by organic synthesis is one of the most important applications of NMR spectroscopy. For this purpose the ^{13}C NMR spectrum plays an important role. Since complex chemical information is encoded in the chemical shift, intensity and multiplicity, these data are suitable to be saved in databases[1,2] and serve for further numerical analysis.[3–9] A broad variety of excellent tools exist that assist the NMR spectroscopists during structure elucidation. However, a fully automated method for elucidating molecular structures from ^{13}C NMR data only has not been realized.

Even for relatively small molecular formulas a huge number of constitutions is theoretically possible. To approach an automated structure elucidation an intelligent structure generator needs to be implemented that uses the experimental ^{13}C NMR spectrum as target function to restrict this huge constitutional space. In contrast to the existing programs Molgen[10] (generates all possible constitutions), CoCon[11] (needs additional 2D NMR connectivity information), and SpecSolv[12] (database-dependent) the approach presented here uses only the molecular formula and ^{13}C NMR chemical shift information and is independent from direct access to databases, since the database is only necessary for training the neural networks but not for predicting the ^{13}C NMR spectra. The exact molecular formula is often known from synthesis or can be experimentally determined by modern high-resolution mass spectrometry. The generated structural space is dynamically determined during the optimization process by a genetic algorithm. The ^{13}C NMR spectra of generated molecules are calculated rapidly and precisely during the optimization process by artificial neural networks.[13] The genetic algorithm starts from a randomly generated set of m molecules for a defined molecular formula (Figure 1). These molecules are created by adding bonds to randomly selected pairs of heavy atoms. Hydrogen atoms are not implicitly considered but assumed to saturate all free valences. This set of molecules undergoes iteratively the processes of selection, recombination, and mutation to minimize the deviation $\Delta(^{13}C)$ of the experimental to the calculated ^{13}C NMR spectrum.

Selection. Artificial neural networks are used to calculate the ^{13}C NMR chemical shifts for each trial chemical structure. The details of the implemented neural networks have been described previously[13] and are therefore summarized only briefly here. The spectra can be predicted for all organic substances that contain exclusively C, H, N, O, P, S, or the halogens. To obtain the spectrum of a molecule the chemical shift of every carbon atom is calculated in an individual run successively. The environment around the carbon atom of interest is subdivided into six spheres. All atoms in these spheres are again separated to belong to one out

Figure 1. Principle of the implemented analysis. The molecular formula and the ^{13}C NMR chemical shifts are obtained from on unknown sample. Using the molecular formula "Genius" creates a set of m random structures. This set of structures is evaluated by comparing the calculated ^{13}C NMR chemical shifts with the experimental data. From the obtained $\Delta(^{13}C)$ values the probabilities are derived for a molecule to be considered for recombination. A new set of m child molecules is created by this process, and the members of this child population undergo a mutation step. This sequence of selection, recombination, and mutation is repeated in an iterative way to optimize the constitution of the molecules until it produces the experimental ^{13}C NMR spectrum with a deviation as low as possible. Ideally, the fittest constitution (structure with the smallest $\Delta(^{13}C)$ value) of the final population is identical with the constitution of the unknown molecule.

of 28 previously defined atom types that consider element number, hybridization, and number of bound hydrogen atoms. The vector containing the number of atoms of every atom type in a particular sphere serves as input for the neural networks. Nine of these 28 atom types describe carbon atoms. For each of these nine types, a separate neural network is trained with the overall number of about 1,300,000 chemical shifts out of the Specinfo database.[1] The average deviation of this method is as low as 1.6 ppm determined for an independent database of about 50,000 chemical shifts. Essential advantages of this method are the fast, exact, and database-independent shift prediction for all organic molecules.[13]

The agreement of the calculated and experimental NMR spectra is assessed by the root-mean-square deviation (RMSD) of the chemical shift deviations of all N carbon atoms of a molecule:

* To whom correspondence should be addressed. Telephone: (206)543-7134. Fax.: (206)685-1792. E-mail: jens@jens-meiler.de.
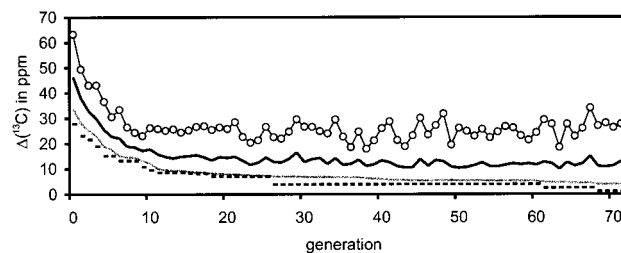† University of Washington.
‡ BASF AG, Ludwigshafen.

$\Delta(^{13}C) \equiv \sqrt{1/N\sum_{i=1}^{N}(\delta_{\text{calc}}^{i}(^{13}C)-\delta_{\text{exp}}^{i}(^{13}C))^2}$. The chemical shifts of the carbon atoms of the generated molecules as well the experimental values are sorted with respect to their size before the comparison is performed. Deviations from the experimental multiplicity (if known) may be included in the RMSD having a **M**ultiplicity **D**eviation **F**actor (MDF). A deviation between experimental and predicted multiplicity is multiplied with this factor and added to the absolute chemical shift deviation: $\Delta(^{13}C) \equiv \sqrt{1/N\sum_{i=1}^{N}(|\delta_{\text{calc}}^{i}(^{13}C)-\delta_{\text{exp}}^{i}(^{13}C)|+\text{MDF}\cdot|M_{\text{calc}}^{i}-M_{\text{exp}}^{i}|)^2}$. The multiplicity of the carbon atoms in the generated structures is computed by analyzing the number of bonded hydrogen atoms. The lower the $\Delta(^{13}C)$ value of a generated structure, the higher its probability to be considered for recombination. The probability for a single molecule j out of a population of m constitutions is given by $p_j = [\Delta_j(^{13}C)]^{-1}/\sum_{i=1}^{m}[\Delta_i(^{13}C)]^{-1}$.

Recombination and Mutation. The numerical vector of all bonds (1 = single, 2 = double, 3 = triple) and nonbonds (0 = nonbonded) between all possible atom–atom pairs is taken as the genetic code of a molecule. Recombination is performed by joining the two vectors representing the genetic code of the two parent molecules. For every position of the newly generated vector of the child molecule one of the two possible states in the two vectors of the parent molecules is randomly considered. The newly created structure is only considered for the child generation if it is chemically reasonable, self-contained, and gives the correct number of hydrogen atoms. A subsequent mutation may be performed by inserting one bond (or increasing bond-type by one) and deleting another bond (decreasing bond-type by one). Optionally, the best l molecules of the parent set of structures are conserved for the child generation without any change. Figure 1 illustrates the algorithm.

The higher the number of heavy atoms and the higher the number of double bond equivalents the more complex is the problem because of the enlarged structural space. Three molecules as well as a small database of 160 structures serve as test examples. All of these molecules are not part of the database used for training the artificial neural networks.

The first example is the amino acid tyrosine ($C_9H_{11}N_1O_3$). With 13 heavy atoms and five double bond equivalents, the problem is relatively small. Nevertheless 2,132,674,846 (!) possible constitutions are calculated by Molgen in 470 min computation time. (All given calculation times are determined on a PII 450 MHz processor equipped with 512 MB memory.) Genius solves this problem in less than 2 min using the experimental ^{13}C NMR spectrum. Sixty-seven generations need to be calculated with 32 molecules ($m = 32$) each. The best eight molecules ($l = 8$) of every generation are automatically considered for the child population. The penalty for wrong multiplicity is set to be MDF = 2 ppm. Figure 2 illustrates the $\Delta(^{13}C)$ values obtained during this experiment. Table 1 summarizes all parameters and results obtained for the three example molecules.

2,3,7,8-Tetrachlorodioxin represents a more complex example. Eighteen heavy atoms and eight double bond equivalents lead to approximately 14.2×10^9 (!) possible constitutions. The high symmetry of the molecule that can be obtained from the ^{13}C NMR spectrum is *not* used to restrict the structure generator only to molecules that meet this symmetry. The Molgen calculation was interrupted after 12 h; about 15% of all constitutions were generated in this time. The computation time necessary for all constitutions would be about 4800 min (more than 3 days). However, to solve the constitution with Molgen an subsequent calculation of the ^{13}C NMR chemical shifts for all suggested constitutions would be
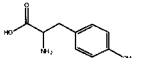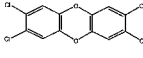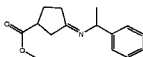

Figure 2. Development of the $\Delta(^{13}C)$ value during the evolution of the tyrosine test population. The average $\Delta(^{13}C)$ of all constitutions of the population (black line), the average of the conserved constitutions (gray line), the best individual (black dotted line), and the worst individual (open circles) are given. While the worst individual stays about constant with a $\Delta(^{13}C)$ value of 25 ppm, the other three monitored $\Delta(^{13}C)$ values decrease until the correct solution structure is found in the 67th generation.

necessary. With our neural networks[13] this would last about 400 days (with the assumption that 5000 s^{-1} chemical shifts are calculated). Genius creates the correct solution structure in 82 min. Four parallel populations ($n = 4$) are calculated independently with 64 molecules ($m = 64$) each. The best 32 molecules ($l = 32$) of a parent population are conserved for the child population. The parallel handling of small sets of individuals in a genetic algorithm is known to accelerate the optimization procedure compared with one large set of individuals. After 270 generations the correct solution occurs for the first time in one out of the four populations.

The molecular formula $C_{15}H_{19}N_1O_2$ of a third example molecule with 18 non-hydrogen atoms covers a structural space of approximately 59.3×10^9 (!) possible constitutions. Molgen is interrupted after about 9% of these constitutions are generated. The genetic algorithm needs 341 generations and 97 min to find the correct solution in one of the four calculations using the same setup as applied for 2,3,7,8-tetrachlorodioxin.

The potential of the approach to be used in fully automated structure elucidation is investigated by testing eight groups containing 20 molecules, respectively. The number of heavy atoms increases from 9 to 16 within these eight groups. The substances are randomly selected from the Specinfo database.[1] A uniform setup of eight parallel populations with 32 molecules is chosen. The eight fittest molecules of the parent population are automatically considered for the child generation. MDF is set to be 1 ppm. For 69% of the structures the correct solution is found. For 31% the calculation is stopped either after 500 generated populations (time limit) or a molecule with a lower $\Delta(^{13}C)$ value than that for the target structure is created (accuracy limit). In both cases the algorithm is considered to have failed. While 85% of all molecules under 15 heavy atoms are predicted correctly, the algorithm fails for 77% of the molecules with 15 and 16 heavy atoms. The size of the populations and the number of generations become too small in these cases. Most of these problems are solved by scaling n and m with respect to the enlarged structural space. However, the accuracy limit also plays an increasing role: for 15% of the molecules containing 15 or 16 heavy atoms, Genius predicts false structures that have a smaller $\Delta(^{13}C)$ value than the correct solution has. The probability of such "false positives" increases with the increasing size of the structural space. The accuracy limit can be addressed by analyzing not only the constitution with the smallest $\Delta(^{13}C)$ value but all generated structures with a $\Delta(^{13}C)$ smaller than the sum of the measurement uncertainty and the neural network prediction error. The time limit is pushed farther and farther by faster processors and parallel computing. Both limits critically depend on the accuracy and velocity of the ^{13}C chemical shift prediction, since it introduces a part of the $\Delta(^{13}C)$ deviation for the correct solution and it is the most time-consuming step in the

Table 1. Molecular Structures, Parameters, and Results Obtained for Some Example Molecules Solved by the Genetic Algorithm Approach

ID	molecule properties								parameters for genetic algorithm				results		
	name	molecular formula	numb. heavy atoms	structure	$\Delta(^{13}C)$ (ppm) [a]	number possible structures [b]	calcu-lation time [c] (min)	struct./ time [d] (min^{-1})	n [e]	m [f]	l [g]	MDF (ppm) [h]	numb. steps [i]	calcu-lation time [j] (min)	struct./ time [k] (min^{-1})
1	tyrosine	$C_9H_{11}NO_3$	13		1.10	$\approx 2.13 \cdot 10^9$	≈ 462	$\approx 4.6 \cdot 10^6$	1	32	8	2	67	2	1237
2	2,3,7,8-Tetra-chlordioxin	$C_{12}H_4O_2Cl_4$	18		1.60	$\approx 14.24 \cdot 10^9$	≈ 4856	$\approx 2.9 \cdot 10^6$	4	64	32	1	270	82	843
3		$C_{15}H_{19}N_1O_2$	18		1.17	$\approx 59.33 \cdot 10^9$	≈ 23522	$\approx 2.5 \cdot 10^6$	4	64	32	1	341	97	900

[a] $\Delta(^{13}C)$ (ppm) value for comparing the experimental data with the NMR spectrum calculated for the correct solution. [b] Total number of possible constitutions generated by Molgen. [c] Calculation time for the generation of all structures by Molgen. [d] Number of generated structures per minute by Molgen (without calculation of the ^{13}C NMR spectra). [e] n: number of parallel calculated populations. [f] m: number of individuals in the populations. [g] l: number of best ranked (small $\Delta(^{13}C)$ value) individuals in the parent generation that are conserved for the new child generation. [h] **M**ultiplicity **D**eviation **F**actor defines the penalty added to $\Delta(^{13}C)$ for a wrong multiplicity of a ^{13}C carbon signal in a generated structure. [i] Total number of steps until the correct solution was found. [j] Total calculation time until the correct solution was found. [k] Number of generated structures per minute (with calculation of the ^{13}C NMR spectra).

algorithm. The accuracy and speed of neural network chemical shift prediction make such an analysis possible for the first time. An additional introduction of fragments that are forbidden (e.g., non-stable structural fragments → bad list) or that have to be used (e.g., known from syntheses → good list) do restrict the accessible structural space further. Such bad and good lists will increase the time and the accuracy limit and are therefore capable of increasing the molecular formulas solvable with this algorithm.

The potential of this approach is proved by its fast and correct handling of the three examples and the small database. The described approach can be run in a highly automated procedure as a first step of structure elucidation. In a second step the algorithm can assist the NMR spectroscopist to analyze the more complex problems which are not solved during the first automated cycle. The approach is limited by the size of the structural space that has to be searched for two reasons: necessary computation time as well as the quality of ^{13}C NMR chemical shift prediction and measurement. The introduction of known (good list) or forbidden (bad list) structural fragments makes it flexible for the use of additional experimental results (e.g., from synthesis or 2D NMR experiments). In combination with this information the size of solvable structures can be increased. The need to know the exact molecular formula could also be circumvented by varying the number of heteroatoms iteratively or by modifying the mutation operator to allow introduction and deletion of heteroatoms. The C++ based computer program "Genius" should become a helpful tool to assist structure elucidation of organic molecules.[14]

References

(1) *SpecInfo Database*; Chemical Concepts: Karlsruhe, 2001.
(2) Robien, W. *Nachr. Chem. Technol. Lab.* **1998**, *46*, 74−77.
(3) Clerc, J.-T.; Sommerauer, H. *Anal. Chim. Acta* **1977**, *95*, 33−40.
(4) Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A. *Carbon-13 NMR Spectral Data*; Verlag Chemie: Weinheim, 1981.
(5) Fürst, A.; Pretsch, E. *Anal. Chim. Acta* **1990**, *229*, 17−25.
(6) Kvasnicka, V.; Sklenak, S.; Pospichal, J. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747.
(7) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.-P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653.
(8) Thomas, S.; Kleinpeter, E. *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504−507.
(9) Svozil, D.; Pospichal, J.; Kvasnicka, V. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924−928.
(10) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. *Anal. Chim. Acta* **1995**, *314*, 141−147.
(11) Lindel, T.; Junker, J.; Köck, M. *J. Mol. Model.* **1997**, *3*, 364−368.
(12) Will, M.; Fachinger, W.; Richert, J. R. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.
(13) Meiler, J.; Will, M.; Meusinger, R. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169−1176.
(14) Meiler, J. http://www.jens-meiler.de, 2001.

JA0109388