CROATICA CHEMICA ACTA
CCACAA **77** (1–2) 213–219 (2004)

ISSN-0011-1643
CCA-2919
Original Scientific Paper

Application of Genetic Algorithms to Structure Elucidation of Halogenated Alkanes Considering the Corresponding ^{13}C NMR Spectra*

Thomas Blenkers and Peter Zinn**

Lehrstuhl für Analytische Chemie, Ruhr-Universität Bochum, D-44780 Bochum, Germany

RECEIVED MAY 5, 2003; REVISED JULY 19, 2003; ACCEPTED AUGUST 20, 2003

A new approach for structure elucidation using genetic algorithms is introduced. In analogy to the genetic programming paradigm developed by Koza, the new concept supports genetic operations on hierarchically coded chemical line notations. The implementation of this concept consists of 5 steps. In the first step, a start population of chemical compounds is randomly generated. As the second step, physical properties of each compound of the population are predicted. The third step is the comparison of each individual property with the observed property of an unknown compound, resulting in the calculation of the fitness value for each generated compound. Depending on the fitness values, the candidates for the next generation are selected by a spinning wheel procedure during the fourth step. In the last step, these candidates are rearranged by genetic mutation and crossover to form the next generation. Steps 2 to 5 of the described procedure are repeated until the spectrum of one candidate is almost equal to the spectrum of the unknown compound within acceptable tolerances. The introduced concept was verified for halogenated alkanes.

Key words
genetic algorithms
structure elucidation
^{13}C NMR spectra

INTRODUCTION

Since the first publications in 1990, the interest in applications of genetic algorithms in chemistry or in closely related sciences has dramatically increased. Today, there are nearly 4000 papers published, with a yearly growth of more than 600 contributions. An overview of the varying application fields of genetic algorithms in chemistry was given by Leardi.[1]

Different approaches have been used to apply genetic algorithms in structure elucidation.[2,3,4] The objects of these investigations are the protein fold prediction and the pharmacophore elucidation, *etc*. Solving molecular crystal structures directly from powder diffraction data using genetic algorithms was reviewed recently.[5] An application of genetic algorithms to structure elucidation in a more general way was given by Meiler.[6,7] This approach requires the measured ^{13}C NMR spectrum of the unknown compound and its experimentally determined molecular gross formula. Depending on the molecular formula, the space for searching the corresponding constitution increases rapidly. Genetic algorithms and ^{13}C NMR spectrum prediction by neural networks are demonstrated as tools to find the constitution corresponding to the measured spectrum.

---

* Dedicated to Professor Nenad Trinajstić on the occasion of his 65th birthday.

** Author to whom correspondence should be addressed. (E-mail: Peter.Zinn@ruhr-uni-bochum.de)

The concept of structure elucidation in the present paper also focuses on ^{13}C NMR spectra and genetic algorithms. As first described in 1995,[8] the present method abstains from the molecular formula and allows structure elucidation without *a priori* structural knowledge.

GENERAL CONCEPT AND IMPLEMENTATION STRATEGY

When including genetic algorithms into the structure elucidation procedure the most important difference from the classical elucidation circuit is that, instead of only one actual structure proposition, a whole population of possible chemical structures has to be considered. Each time the population passes through the elucidation circuit, the individual structures are modified by genetic operations. Figure 1 shows the general concept of structure elucidation using genetic algorithms.

The starting point of the genetic algorithm is the generation of an initial population of chemical structures. In generating an initial population the main problem is to find a structure notation that allows genetic modifications of the individual structures *e.g.* mutation and crossover. Genetic algorithms can conveniently process binary coded individuals. These individuals are often represented as bit strings of fixed length or as the corresponding real numbers. On the other hand, typical computer-oriented molecular codes[9] are far from bit strings or real number coding. The three most widespread structure notations are the connection table, the different types of line notations, and the adjacency matrix.[10] Transformation of one of these representations into a bit string or a real number is difficult and results in data structures unsuitable for genetic operations. *E.g.,* transformation of the adjacency matrix to a bit string might succeed in linking the single lines or columns of the matrix one after the other. Genetic modifications of such bit strings could result in structures containing atoms with wrong valencies.

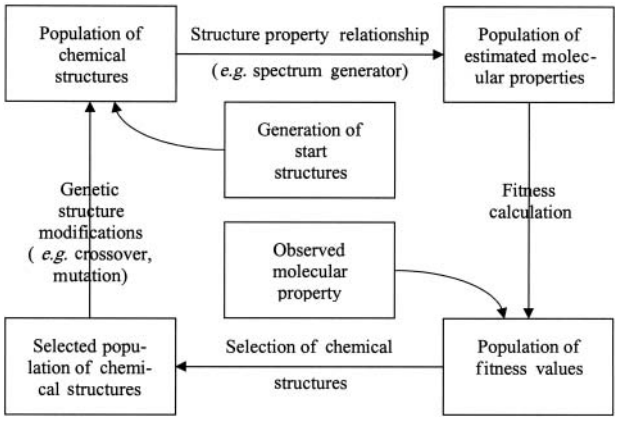

Figure 1. Overview of the iterative structure elucidation strategy applying genetic algorithms. Instead of one chemical compound, whole generations of substances are passing through the circuit.

To allow only genetic modifications that result in correct chemical structures, a system of corresponding constraints and modification rules seems to be necessary. This complication slows down the genetic algorithm and decreases practicability.

A completely different way of coding the individuals of a genetic algorithm was introduced by J. Koza.[11] Koza's Genetic Programming is a technique for finding the best mathematical formula in order to solve a given problem. The principal idea is to code formulas in a tree-like hierarchical way and to modify them by genetic operations from generation to generation until an optimal result is reached. Within a tree, sub-trees can be cut off and substituted by sub-trees from other trees in the sense of a genetic crossover. Point mutations can be also performed easily. In the following chapter, we will discuss in detail the transformation of this coding to chemical structures.

As the second step of the genetic structure elucidation, the prediction of a physical property of each individual chemical structure is to be performed. Physical properties, *e.g.* spectral, diffraction, chromatographic, thermodynamic or other data, can be used if a well defined quantitative structure property relationship (QSPR) is known. Besides a single QSPR, the prediction of different properties or a combined QSPR is also of interest and may improve the accuracy of the results of the genetic structure elucidation. Especially in cases when single QPR's are of less predictive quality, a combination may be successful. A set of property values corresponding to each individual structure of the actual population is the result of this step.

As the third step, the fitness value for each individual has to be calculated, including the comparison between the observed and the estimated property. Typical fitness functions are error measures such as the root mean square, mean absolute error, *etc*. In the case of combined properties, different fitness measures can be taken into account with respect to the reliability of the estimated properties. If the fitness value of an individual or the mean fitness of the complete population becomes better than the threshold value, the genetic algorithm stops; otherwise, the algorithm is continued with the next step.

The fourth step is responsible for the selection of the individuals for the next generation. Depending on the fitness values, a random procedure decides which individuals or pairs of individuals are selected for mutation or crossover operations. A typical selection procedure is the spinning wheel method[12] with circle sectors proportional to the fitness values of the individuals.

In the fifth step, genetic operations applied to hierarchically notated chemical structures are performed. Concerning the genetic crossover, randomly chosen sub-trees of two individuals are exchanged. Concerning the genetic mutation, a randomly chosen sub-tree of a single in-

dividual is eliminated and substituted by a randomly generated new sub-tree to form an individual for the next generation. After the fifth step, the first elucidation circuit is finished and a new generation of chemical structures has been built to pass the next circuit, and so on.

Because the genetic operations as well as the generation of the start population have been verified as symbolic programming techniques instead of conventional numerical methods of the other steps, these operations are described in more detail in the next two chapters.

GENERATION OF HIERARCHICALLY CODED CHEMICAL STRUCTURES

In a previous paper, we have introduced the basic list data types for performing list operations on chemical graphs.[13] Among these data types, we have given an example of a tree – like hierarchical notation of a chemical structure. This hierarchical structure is based on a line notation that is a near SMILES[14] implementation in the programming language LISP.[15] As an example, Figure 2 gives the line notation of 1-bromo-2-chlorobutane.

In order to generate a hierarchical representation needed in the genetic algorithm, a root atom has to be determined. Then, corresponding to the root atom, branches and sub-branches are added, characterized by additional parentheses. The hierarchical representation of the example molecule is also shown in Figure 2. Obviously, the hierarchical notation as well as the line notation of a chemical compound is not unique. Even if the root atom is determined, different notations are possible dependending on the order of branches and sub-branches. With respect to the genetic algorithm, uniqueness of the notation is not necessary provided that all possible notations of a compound result in the same value during the property estimation of step 2 of the algorithm.

In order to discuss the principle of the generation of hierarchically represented chemical structures, we have chosen the class of chlorinated and brominated alkanes as example compounds. In this context, it is important that the corresponding compound space contains all possible chlorinated, brominated, and mixed halogenated congeners as well as pure alkanes as potential candidates. In analogy to Koza's Gentic Programming, we classify all skeleton atoms of a chemical structure into terminal or node atoms. Corresponding to the example class of halogenated alkanes, the following terminal and node sets result:

Terminal set: CH_3, Cl, Br
Node set: C, CH, CH_2

The generation of a tree-like chemical structure starts with the selection of the root atom as a randomly chosen atom from the union of terminal and node sets. The generation procedure stops if the selected atom is a member of the terminal set. It continues with a recursive procedure if the actual atom belongs to the node set. The number and kind of the elements of the terminal and node sets are responsible for the magnitude of the generated trees. For example, multiple occurrence of CH_3-groups in the terminal set reduces the magnitude of the trees while multiple occurrence of node set atoms expands it. Also, the degree of halogenation of the generated compounds can be influenced by the ratio between halogens and other atoms.

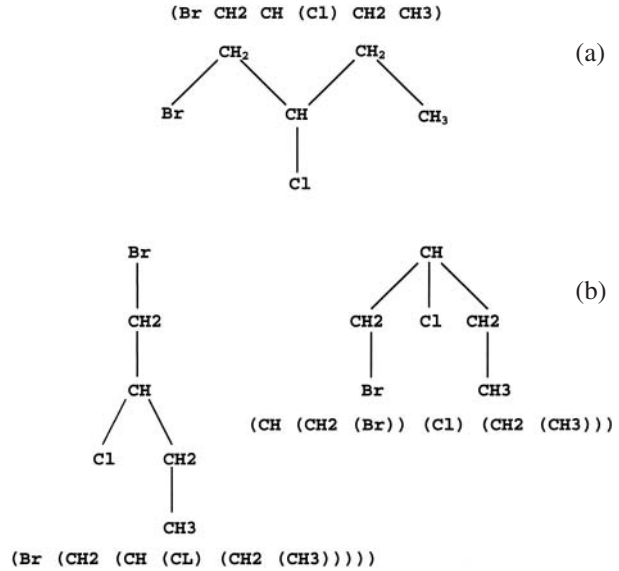

Figure 2. Example of a hierarchical coded chemical line notation. Because it is not unique several notations are belonging to the same molecule. A corresponding property prediction has to result in the same value independent of the hierarchical notation of the molecule. (a) Line notation of 1-bromo-2-chlorobutane. (b) Hierarchical representations of 1-bromo-2-chlorobutane.

GENETIC OPERATIONS APPLIED TO HIERARCHICAL CODED CHEMICAL STRUCTURES

The most important genetic operation is the crossover of two molecules. Figure 3 illustrates the crossover procedure with the example molecules 1,2-dichlorohexane and 3-methylheptane. First, the random selection is performed of one atom of each molecule as root atoms for the corresponding sub-trees. In the next step, the sub-trees defined by their root positions are extracted from the trees. Finally, the sub-trees are swapped, rearranging the two new molecules 1,2-dichloro-5-methylheptane and n-hexane as shown in the example.

In cases of large compound spaces, genetic mutation is of increasing significance. Especially if the compound population has settled after some elucidation cycles near a local maximum of fitness, mutations can increase the structural variability of the population in order to find the global maximum. Mutation is done by either cutting some atoms at the end of a (side-)chain or introducing a

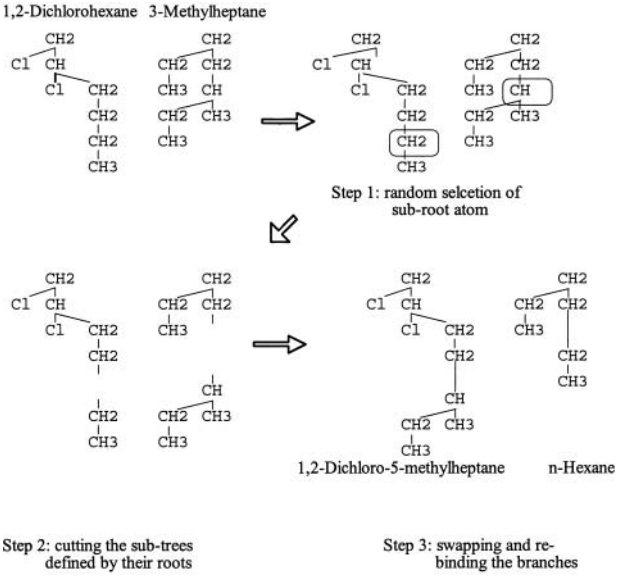

Figure 3. Single steps of the genetic crossover applied to hierarchically notated chemical compounds.

randomly generated side-chain at a random position. In the latter case, the mutation can also be understood as applying the first two steps of the described crossover operation to one molecule only, followed by the insertion of a randomly generated side-chain. The complete program code of the genetic operations can be found in Blenkers.[8]

^{13}C NMR SPECTRUM GENERATOR

When realizing the introduced concept of the quantitative structure property relationship, a spectrum generator must be available in order to perform the fitness calculation. Because of some experiences with the ^{13}C NMR spectrum generation[16] and the corresponding straightforward implementation of increment rules,[17] such a spectrum generator was chosen. Chlorinated and brominated alkanes were selected as the investigated compound group. As a consequence of genetic crossover operations, besides halogenated compounds, also ^{13}C shifts of pure alkanes must be predictable by the spectrum generator.

Developing the spectrum generator, the ^{13}C shifts of compounds including 1–12 carbon atoms and 1–5 chlorine or bromine atoms were extracted from the SPECINFO data base.[18] In order to guarantee high homogeneity of the measurement conditions, the retained data set includes only spectra meeting the following criteria:

– Deutero-chloroforme was used as solvent.
– Measurement temperature ranged from 17 to 37 °C.
– TMS or $CDCl_3$ were used as standards.

Under these conditions, the final data set contained 118 chlorinated and brominated acyclic alkanes, including 651 ^{13}C shifts. Building up a prediction model for high-precision prediction of chemical shifts, the model increments were derived by multiple linear regression:

$$\delta_{ipso} = \delta_{\text{Methan}} + \sum_{m,\text{X}} N_{m,\text{X}} \cdot S_{m,\text{X}}$$

Increments $S_{m,\text{X}}$ are the resulting regression coefficients and $N_{m,\text{X}}$ are the corresponding descriptors. Descriptors are the counts of different atoms or groups (X = C, CH, CH2, CH3, Cl, Br) in different spheres (m = α, β, γ) around the observed *ipso*-carbon atom. All the considered descriptors and corresponding increments are given in Table I. Additionally, the standard errors and the frequencies of occurrence of the descriptors ensure that the incorporated 29 increments are significant. In accordance with the NMR theory, we found significant influences on the chemical shift only in the α, β, and γ spheres, with decreasing importance. Furthermore, squared descriptors and

TABLE I. Descriptors of chlorinated and brominated acyclic alkanes for the prediction of ^{13}C NMR shifts[(a)]

Descriptor	Frequency	Coefficient	Standard error
$N_{\alpha CH3}$	115	32.69	0.72
$N_{\alpha CR}$	637	27.62	0.72
$N_{\alpha Cl}$	133	26.74	0.40
$N_{\alpha Br}$	67	17.20	0.55
$N_{\beta CH3}$	163	14.81	0.81
$N_{\beta CH2}$	458	11.33	1.05
$N_{\beta CH}$	90	8.17	1.10
$N_{\beta Cq}$	39	8.51	1.57
$N_{\beta Hal}$	241	–3.64	0.40
$N_{\gamma C}$	489	4.22	0.51
$N_{\gamma Br}$	66	–2.43	0.46
$N_{\alpha H}\ N_{\gamma C}$	471	–2.09	0.19
$N_{\alpha H}\ N_{\gamma Cl}$	118	–1.41	0.18
$N_{\alpha CH3}\ N_{\gamma Cl}$	21	1.31	0.26
$N_{\alpha CH3}\ N_{\gamma Br}$	9	2.86	0.55
$N^2_{\beta H}$	578	–0.68	0.02
$N^2_{\delta H}$	477	–0.08	0.01
$N^2_{\alpha Br}$	67	–3.78	0.30
$N^2_{\alpha Cl}$	133	–0.49	0.13
$N_{\alpha H}\ N_{\beta CH3}$	159	–2.69	0.16
$N_{\alpha H}\ N_{\beta CH2}$	448	–1.61	0.21
$N_{\alpha H}\ N_{\beta Cq}$	32	1.00	0.42
$N_{\alpha H}\ N_{\beta Hal}$	110	3.18	0.10
$N_{\alpha H}\ N_{\beta H}$	540	–1.59	0.07
$N_{\alpha CH3}\ N_{\beta C}$	93	–7.28	0.36
$N_{\alpha CH2}\ N_{\beta C}$	427	–3.74	0.23
$N_{\alpha CH}\ N_{\beta C}$	134	–4.61	0.25
$N_{\alpha Cq}\ N_{\beta C}$	90	–4.48	0.32
$N_{\alpha Cl}\ N_{\beta C}$	229	–2.93	0.27

[(a)] Columns 2–4: frequency of descriptors occurrence in the data set, the corresponding increments and standard errors (in ppm).

different types of combined descriptors were of significant influence in modeling ^{13}C shifts. An illustration of the robustness of the derived increment model is given in Figure 4. Plotting the predicted *versus* the observed chemical shifts, a standard error of 1.65 was calculated. This result is acceptable if one keeps in mind that the ^{13}C shifts included in the SPECINFO data base and the model development stem from different origins. Calculation of the intensities of the NMR peaks was not considered. All peaks were treated as being of equal height.

FITNESS CALCULATION

The spectrum generator is the basis for the calculation of the fitness value between the spectrum of an intermediately generated individual molecule and the observed target spectrum. Fitness should express the similarity between the two spectra. Large similarity should be expressed by a high numerical value and a great difference should give a small value. A widely used value for the comparison of two spectra is the sum of the squares of errors (differences), SSE. Because the SSE is countercurrent to fitness, it is transformed to a fitness conformable value:

$$\text{fitness} = 1 / (\text{SSE} + 1)$$

This fitness definition ensures that a difference of 0 can also be treated. The fitness value equal to 1, which means zero difference between the two spectra, is used as an aborting criterion for the genetic algorithm. For easier interpretation of the results, the reciprocal value 1/fitness called »malus« or »worse value« is also used.

Direct comparison of the observed and predicted spectra is not suitable because in many cases no direct association between the observed and predicted NMR signal is possible. Therefore, the range of the shifts is restricted from 0 to 100 ppm and is divided in a fixed number of sectors, typically 100 (one sector = 1 ppm). Whenever a chemical shift is calculated, the numerical value of the corresponding sector is increased by one. At the end of the calculation, the spectrum is represented by a vector of numbers, mostly 0's, some 1's and a few higher numbers for peaks of nearly isochronic atoms in the same sector. Figure 5 shows the described spectral representation in detail.

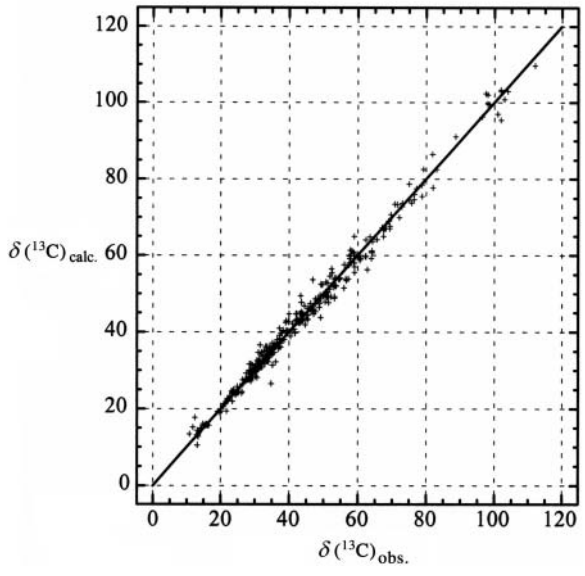

Figure 4. Observed *versus* predicted chemical shifts resulting from the multiple linear regression model including the descriptors from Table I. The standard error of the model is SE = 1.65 ppm.

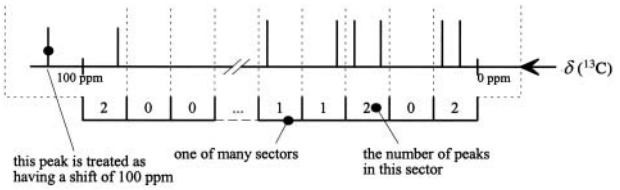

Figure 5. Transformation of the ^{13}C NMR spectra from the SPECINFO database into data vectors with 100 elements in the range of 0 to 100 ppm. The resulting example vector will be (2 0 0... 1 1 2 0 2).

A sectorwise comparison between the observed and the corresponding predicted peaks has the same influence on the fitness value regardless of the difference from the expected sector. Therefore, a kind of fuzzy comparison is involved, increasing the half width of each peak over several sectors for the observed and predicted spectra. After the transformation, it is much more likely that a predicted and an observed peak overlap within their broadened area if their centers are next to each other. Figure 6 gives an example of the fuzzification included in the complete operation of the fitness calculation.

SYSTEM VALIDATION AND RESULTS

The elucidation procedure was implemented in the programming language LISP[19] running under the Suse LINUX 6.0 operating system. Although LISP offers a compact programming style, it is time consuming. Elucidation of one component running on a Pentium2 platform with 110 MHz takes about 0.5 hours. More powerful computer equipment would decrease the computation time to an acceptable length.

The determination of the terminal and node sets of the skeleton atoms such as C, CH, CH2, CH3, Cl, and Br is orientated to the frequency of the skeleton atoms in the data set in such a way that the magnitude of the generated molecules is almost comparable with those of the data set.

Genetic modification of individuals is conveniently restricted to crossover probabilities of high percentages and mutation probabilities of low values. Besides both basic genetic operations, we additionaly considered an elitism, which means that the fittest individuals were kept

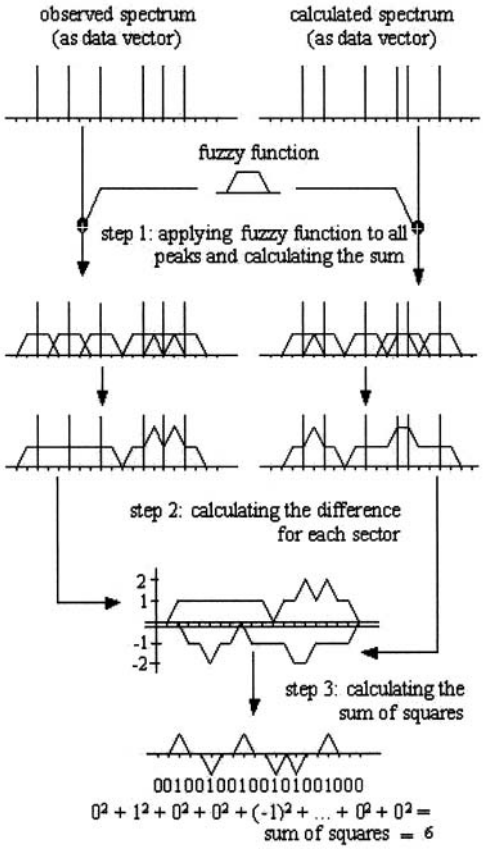

Figure 6. Example of the fitness calculation including a fuzzification step and the building of the sum of squares of spectral differences.

from one generation to the other. In several test runs this strategy was found to be successful.

In order to validate the introduced new approach, an attempt was made to find the structure of each of the 118 molecules of the data set. The validation parameters were kept unchanged during the validation runs:

– Terminal set: (CH3 CH3 CH3 Cl Br),
– Node set: (C CH CH2 CH2),
– Evolution time: 50 generations,
– Population size: 25 individuals,
– Elitism: 1 molecule,
– Probability of crossover: 90 %,
– Probability of mutation: 30 %,
– Fuzzy function: (1 1 1 2 3 3 3 2 1 1 1),
– Number of slots: 100 (1 slot = 1 ppm).

Under the conditions of the parameter set validation, Figure 7 gives the first impression of the results. The already defined worse value or malus of the best molecule at the end of a test run is shown *versus* the indexes of the molecules in the data set. The bars pointing downwards indicate runs where the correct molecules could not be found. With only one exception, such molecules have a worse value of more than 50. On the other hand, these values can be obtained for some other molecules as well. Therefore, it was proven to be better for the interpretation of the results to use the quotient of the »worse value of the best molecule found« / »worse value of the expected molecule«. According to this quotient, we obtained three different groups of molecules, as illustrated in Figure 8:

– Group one: molecules found with the GA, quotient = 1 (53 molecules).
– Group two: molecules that might be found, quotient > 1 (43 molecules).
– Group three: molecules that cannot be found, quotient < 1 (22 molecules).

The number of molecules in group one is of high significance compared to the number that would be expected only by pure chance retrieving without genetic modifications. Consequently, it is demonstrated that chemical structures could be predicted applying genetic algorithms without any structural *a priori* knowledge.

Prediction improvements can be made easily by initiating the GA with the generation of other start popula-

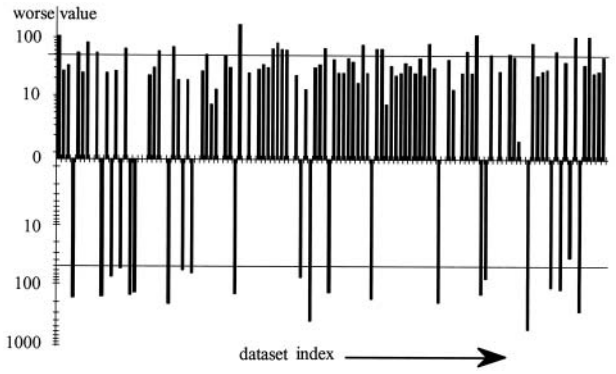

Figure 7. Worse values (= reciprocal fitness values) of the 118 molecules used in the data set for calculation of the prediction system. Lines pointing down indicate the molecules of group 3 and still have positive values. The two horizontal lines represent a worse value of 50, which mostly separates good from bad solutions.

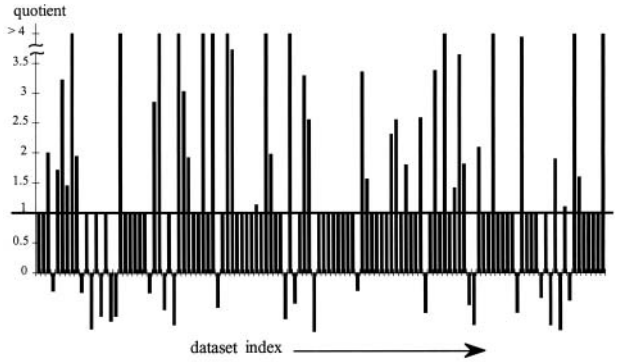

Figure 8. If the quotient »worse value of the best molecule found« / »worse value of the target molecule« is less than 1, the found molecule has a lower worse value than the expected molecule would have. The expected molecule cannot be found in such cases. The horizontal line represents a quotient of one. Bars lying on this line indicate molecules that were directly found using the given system parameter set.

tions or by modifing genetic operations. For that purpose, after changing the system parameters in other test runs, all the molecules in group two could also be identified correctly.

The molecules in group three could not be found, even if the GA was modified. The size of this group may be decreased by increasing the precision of the increment system. A limitation in precision will still remain because the applied increment system is based on pure topological instead of more fundamental geometric models for characterization of the molecules. For example, some molecules of group three are diastereomeres with different shifts from each other. Therefore, the corresponding chemical shifts can only be computed correctly if the geometric arrangement is incorporated into the prediction model. Another limitation may be established in the accuracy of the available spectra in the utilized SPECINFO data base.

CONCLUSIONS

The presented investigation has shown that the prediction of chemical constitutions is possible applying genetic algorithms to NMR data without any *a priori* structural knowledge. Although the investigated compound class is restricted to halogenated alkanes, the approach can be transferred to other substance classes provided a precise property or spectrum generator is available. If the precision of the property predictor is not sufficient, a combination of different properties *e.g.* the intensity and multiplicity of NMR signals or the incorporation of *a priori* structural knowledge can be helpful. In order to expand the approach to cyclic compound classes, a quasi hierarchical structure notation has to be developed. A more difficult problem is the incorporation of geometric isomerism because the described approach uses pure topologically based structures for the hierarchical representation of the molecules. Essential modifications seem to be nessesary, including the corresponding isomerisms and 3D-QSAR approaches.

REFERENCES

1. R. Leardi, *J. Chemometrics* **15** (2001) 559–569.
2. T. Dandekar and P. Argos, *Protein Eng.* **10** (1997) 877–893.
3. G. Jones, P. Wilett, and R. C. Glen, *Comput. Aided Mol. Des.* **9** (1995) 532–549.
4. B. Contreras-Moreiro, P. W. Fitzjohn, and P. A. Bates, *J. Mol. Biol.* **328** (2003) 593–608.
5. W. I. David, K. Shankland, and A. Markvardsen, *Crystallogr. Rev.* **9** (2003) 3–15.
6. J. Meiler and M. Will, *J. Am. Chem. Soc.* **124** (2002) 1868–1869.
7. J. Meiler and M. Will, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1535–1546
8. T. Blenkers, Degree Thesis, Ruhr-Universität, Bochum, 1995.
9. J. V. Knop, W. R. Mueller, K. Szymanski, S. Nikolić, and N. Trinajstić, *Comput. Chem. Graph Theory* (1990) 9–32.
10. A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **25** (1985) 334–343.
11. J. R. Koza, *Genetic Programming*, MIT Press, Cambridge Mass., 1993.
12. D. E. Goldberg, *Genetic Algorithms in Search, Optimisation and Machine Learning.* Addison-Wesley, Reading Mass., 1989.
13. R. Gautzsch and P. Zinn, *J. Chem. Inf. Comput. Sci.* **32** (1992) 541–550.
14. D. Weininger, *J. Chem. Inf. Comput. Sci.* **28** (1988) 31–36.
15. G. L. Steele, *Common Lisp.* Digital Press, 1990.
16. Chr. Duvenbeck, Ph.D. Thesis, Ruhr-Universität Bochum, 1995.
17. D. M. Grant and E. G. Paul, *J. Am. Chem. Soc.* **86** (1964) 2984–2990.
18. SPECINFO Database, Chemical Concepts GmbH, Weinheim, 1995.
19. B. Haible and M. Stoll, *CLISP.* http://clisp.cons.org/.

SAŽETAK

Primjena generičkih algoritama na razjašnjavanje strukture halogeniranih alkana pomoću njihovih ^{13}C NMR spektara

Thomas Blenkers i Peter Zinn

Uveden je novi postupak za razjašnjavanje strukture molekula pomoću genetičkih algoritama. U analogiji s paradigmom genetičkoga programiranja, koju je razvio Koza 1993., nova koncepcija podupire genetičke operacije na hijerarhijski kodiranim kemijskim linearnim notacijama. Implementacija ove koncepcije sastoji se od pet koraka. U prvom se koraku nasumice generira početni skup kemijskih spojeva. U drugom se koraku predviđaju fizikalna svojstva svakoga spoja u skupu. U trećem se koraku uspoređuje svako pojedino svojstvo s izmjerenim svojstvom nepoznatoga spoja što rezultira u računanju vrijednosti podešavanja (*fitness value*) za svaki generirani spoj. Ovisno o vrijednosti podešavanja biraju se kandidati za sljedeći korak pomoću postupka nazvanoga kolovrat (*spinning wheel*). To je četvrti korak, a u petom, posljednjem koraku odabrani kandidati se preslože pomoću genetičke mutacije i formiraju novu generaciju. Postupak je iterativan i ponavlja se sve dok nije spektar jednoga kandidata skoro jednak spektru nepoznatoga spoja unutar prihvatljivih odstupanja. Predloženi je postupak provjeren na halogeniranim alkanima.