Anal. Chem. **2002,** *74,* 80−90

Prediction of ^{1}H NMR Chemical Shifts Using Neural Networks

João Aires-de-Sousa,[†] Markus C. Hemmer,[‡] and Johann Gasteiger*[,‡]

Computer-Chemie-Centrum Institut fur Organische Chemie, Universität Erlangen, Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Secção de Quimica Organica Aplicada, Departamento de Quimica, CQFB, Fac. Cienciase Technologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Counterpropagation neural networks were applied to the fast prediction of ^{1}H NMR chemical shifts of CH_n groups in organic compounds. The training set consisted of 744 examples of protons that were represented by physicochemical, topological, and geometric descriptors. The selection of descriptors was performed by genetic algorithms, and the models obtained were compared to those containing all the descriptors. The best models yielded very good predictions for an independent prediction set of 259 cases (mean absolute error for whole set, 0.25 ppm; mean absolute error for 90% of cases, 0.19 ppm) and for application cases consisting of four natural products recently described. Some stereochemical effects could be correctly predicted. A useful feature of the system resides in its ability to be retrained with a specific data set of compounds if improved predictions for related structures are required.

Fast and accurate predictions of ^{1}H NMR chemical shifts of organic compounds are highly desired for the analysis of combinatorial libraries, for automatic structure elucidation, and for the interpretation of spectra by chemists and spectroscopists.

NMR chemical shifts can be predicted by ab initio calculations,[1] and much effort has been devoted to this endeavor. However, the computation times required for such an approach are considerable, and for most organic compounds, the accuracy obtained is not better than faster empirical methods.

The available empirical methods[2] are based on a vast amount of data collected particularly from ^{1}H and ^{13}C NMR spectra of stable organic compounds and have followed two main approaches. In the first one,[3] a database of atom structural environments (usually represented by HOSE codes[4]) is searched and the prediction of the chemical shift is based on the chemical shifts of the most similar atoms found. Such a method is used, for example, in the commercial ACD/I-Lab[5] and Chemical Concepts SpecInfo[6] packages. In the second approach,[7] predictions are based on tabulated chemical shifts for classes of structures and corrected with additive contributions from neighboring functional groups or substructures. Several tables have been compiled for different types of nuclei.[7] Examples of such implementations include ChemDraw (CambridgeSoft)[8] and TopNMR (Upstream).[9] It has been claimed that the incremental methods can predict the ^{1}H NMR shifts of ∼90% of all CH_n groups with a mean deviation between 0.2 and 0.3 ppm.[9]

For the prediction of ^{13}C NMR chemical shifts, a third approach using neural networks or linear regression modeling has been tried which used topological, physicochemical, or geometric parameters as independent variables.[10] Predictions have been reported that are at least in the same range of accuracy as those obtained by the other methods.[10h−j] A trend to obtain better results with neural networks as compared to regression methods could be observed.[10f,h,i]

Rigorous comparison of the methods is however risky, as the results are strongly dependent on the data used. It seems more likely that each approach is complementary to the others. A user will be in a safer position if predictions by different methods are available.

Estimation of ^{1}H NMR chemical shifts is usually more problematic than the prediction of ^{13}C chemical shifts due to the strong influence of experimental conditions and of structural effects that are difficult to model. This is particularly true for the influence of stereochemistry and of 3D geometry.

*To whom correspondence should be addressed. Telephone: 0049-9131-85-6570. Fax: 0049-9131-85-6566. E-mail: Gasteiger@chemie.uni-erlangen.de.

[†] Universidade Nova de Lisboa.

[‡] Universität Erlangen-Nürnberg.

(1) For examples. see: (a) Czernek, J.; Sklenár, V. *J. Phys. Chem. A* **1999**, *103*, 4089−4093. (b) Barfield, M.; Fagerness, P. *J. Am. Chem. Soc.* **1997**, *119*, 8699−8711.

(2) Williams, A. *Curr. Opin. Drug Discovery Dev.* **2000**, (June) (available from ACD web site at http://www.acdlabs.com/publish/publ_pres.html).

(3) Schütz V.; Purtuc V.; Felsinger S.; Robien W. *Fresenius J. Anal. Chem.* **1997**, *359*, 33−41.

(4) Bremser W. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(5) ACD/Ilab Interactive Laboratory, http://www.acdlabs.com/ilab. Calculations performed in May 2001.

(6) Chemical Concepts, http://www.chemicalconcepts.com.

(7) (a) Schaller R. B.; Pretsch E. *Anal. Chim. Acta* **1994**, *290*, 295−302. (b) Fürst, A.; Pretsch, E. *Anal. Chim. Acta* **1990**, *229*, 17−25.

(8) ChemDraw Ultra 4.5, CambridgeSoft, http://www.cambridgesoft.com.

(9) Upstream, http://www.upstream.ch.

(10) (a) Kvasnièka, V.; Sklenak, S. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747. (b) Svozil, D.; Pospíchal, J.; Kvasnièka, V. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924−928. (c) Ball, J. W.; Anker, L. S.; Jurs, P. C. *Anal. Chem.* **1991**, *63*, 2435−2442. (d) Ball, J. W.; Jurs, P. C. *Anal. Chem.* **1993**, *65*, 505−512. (e) Anker, L. S.; Jurs, P. C. *Anal. Chem.* **1992**, *64*, 1157−1164. (f) Clouser, D. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168−172. (g) Doucet, J. P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324. (h) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653. (i) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 587−598. (j) Meiler, J.; Meusinger, R.; Will, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169−1176.

10.1021/ac010737m CCC: $22.00
Published on Web 12/01/2001

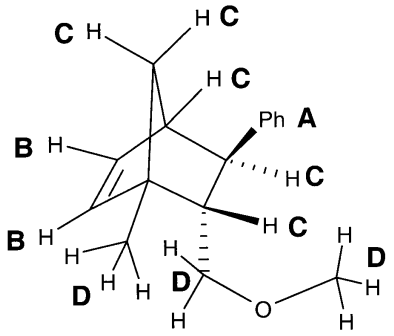


Figure 1. Example of classification of protons from a molecule according to the four classes defined in text (A, aromatic; B, nonaromatic π; C, rigid aliphatic; D, nonrigid aliphatic).

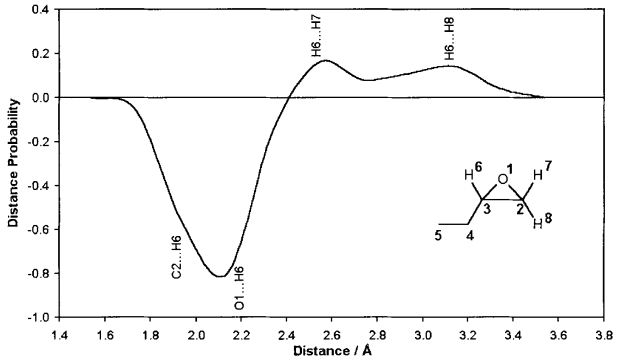


Figure 2. Radial distribution function $g_H(r)$ for proton H-6 and indications of the distances contributing to each peak.

Only in specific situations,[7a,11] have stereochemistry and 3D effects been addressed in the context of empirical ^{1}H NMR chemical shift prediction.[12] The reason for this may reside in the lack of stereochemical labeling (and assignments for diastereotopic protons) in the available databases of spectra,[7b] but also in the lack of suitable representations for the 3D environment of hydrogen nuclei.[13]

In this article, a strategy is described for the fast estimation of NMR chemical shifts of CH_n protons, which is based on automatic knowledge acquisition from a set of experimental examples. The system was designed in order that learning of 3D effects is possible. The relationship between protons in defined molecular structures and the corresponding ^{1}H NMR chemical shift was established by counterpropagation neural networks (CPG NN),[14] which used descriptors for hydrogen atoms in organic structures as input and the chemical shift of the corresponding proton as output.

To ensure robustness and generality, various types of descriptors were used, namely, topological and physicochemical descriptors. Geometric descriptors were added in some situations to account for stereochemistry and 3D effects.

Different conformations usually contribute to the observed chemical shifts when a spectrum is measured in solution. This situation makes the definition of the 3D environment of a proton rather difficult. However, when a given proton belongs to a rigid substructure, a single conformation can more safely be considered and the geometry of the environment be defined. In this work, we took that approximation to characterize the 3D environment and to calculate geometric descriptors for protons of rigid substructures. The developed geometric descriptors were based on radial distribution functions (RDF) that have been proposed for fixed-length representation of 3D molecular structures.[15]

Four different types of protons were treated separately: protons belonging to aromatic systems, protons belonging to π nonaromatic systems, protons belonging to rigid aliphatic substructures, and protons belonging to nonrigid aliphatic substructures. The CPG NNs were trained with training sets and tested with independent prediction sets.

It is common to use selected subsets of descriptors, instead of all possible descriptors, to develop more compact and robust models.[16] We chose genetic algorithms[17] to select subsets of descriptors for each of the four classes of protons and compared the results with those obtained by the models that used all the descriptors. Four examples of natural products recently discovered were chosen to test the best models and to compare the predictions of chemical shifts with those provided by two commercial programs.

METHODOLOGY

Data Sets and Classification of Protons. A training set of 744 ^{1}H NMR chemical shifts for protons from 120 molecular structures (Appendix A) and a prediction set with 259 chemical shifts from 31 molecules (Appendix B) were collected from the literature.[18] Only data from spectra obtained in $CDCl_3$ were

(11) For recent works, see: (a) Martin, N. H.; Allen, N. W., III; Brown, J. D.; Ingrassia, S. T.; Minga, E. K. *J. Mol. Graphics Modell.* **2000**, *18*, 1−6. (b) Martin, N. H.; Allen, N. W., III; Moore, J. C. *J. Mol. Graphics Modell.* **2000**, *18*, 242−246.

(12) For empirical predictions of ^{13}C NMR chemical shifts accounting for 3D influences see refs 3 and 10c and e and the following: (a) Bernassau, J. M.; Fetizon, M.; Maia, E. R. *J. Phys. Chem.* **1986**, *90*, 6129−6134. (b) Hönig, H. *Magn. Reson. Chem.* **1996**, *34*, 395−415.

(13) Canonical representation of molecular 3D structures including stereochemistry has been proposed in: Satoh, H.; Koshino, H.; Funatsu, K.; Nakata, T. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 622−630. For extensions of the HOSE code for inclusion of stereochemical information in the context of ^{13}C NMR spectroscopy see ref 3 and: Gray, N. A. B.; Nourse, J. G.; Crandell, C. W.; Smith, D. H.; Djerassi, C. *Org. Magn. Reson.* **1981**, *15*, 375−389.

(14) For detailed description of neural networks, see: (a) Gasteiger, J.; Zupan, J. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527; *Angew. Chem.* **1993**, *105*, 510−536. (b) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999. (c) Hecht-Nielsen, R. *Appl. Optics* **1987**, *26*, 4979−4984.

(15) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectrosc.* **1999**, *19*, 151−164.

(16) (a) Winkler, D. A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430 and references therein. (b) Kubinyi, H. *J. Chemom.* **1999**, *10*, 119−133.

(17) Jones, G. Genetic and Evolutionary Algorithms. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds; John Wiley & Sons: Chichester, U.K., 1998; pp 1127−1136.

(18) (a) Ahmed, A. A.; El-Razek, M. H. A.; Mostafa, E. A. A.; Williams, H. J.; Scott, A. I.; Reibenspies, J. H.; Mabry, T. J. *J. Nat. Prod.* **1996**, *59*, 1171−1173. (b) Tuntiwachwuttikul, P.; Boonrasri, N.; Bremner, J. B.; Taylor, W. C. *Phytochemistry* **1999**, *52*, 1335−1340. (c) Chen, K.; Zhang, Y.-L.; Li, Z.-L.; Shi, Q.; Poon, C.-D.; Tang, R.-J.; McPhail, A. T.; Lee, K.-H. *J. Nat. Prod.* **1996**, 59, 1200−1202. (d) de la Fuente, G.; Gavin, J. A.; Acosta, R. D.; Sanchez-Fernando, F. *Phytochemistry* **1993**, *34*, 553−558. (e) Duran-Patron, R.; O'Hagan, D.; Hamilton, J. T. G.; Wong, C. W. *Phytochemistry* **2000**, *53*, 777−784. (f) Lien, T. P.; Kamperdick, C.; Sung, T. V.; Adam, G.; Ripperger, H. *Phytochemistry* **1998**, *49*, 1797−1799. (g) Sy, L. K.; Brown, G. D. *Phytochemistry* **1998**, *49*, 1715−1717. (h) Murakami, R.; Shi, Q.; Oritani, T. *Phytochemistry* **1999**, *52*, 1577−1580. (i) Tazaki, H.; Iwasaki, T.; Nakasuga, I.; Kobayashi, K.; Koshino, H.; Tanaka, M.; Nabeta, K. *Phytochemistry* **1999**, *52*, 1427−1430. (j) Chen, I.-S.; Chen, T.-L.; Lin, W.-Y.; Tsai, I.-L.; Chen, Y.-C. *Phytochemistry* **1999**, *52*, 357−360. (k) Montagnac, A.; Martin, M.-T.; Debitus, C.; Païs, M. *J. Nat. Prod.* **1996**, *59*, 866−868. (l) Takano, I.; Yasuda, I.; Nishijima, M.; Hitotsuyanagi, Y.; Takeya, K.; Itokawa,

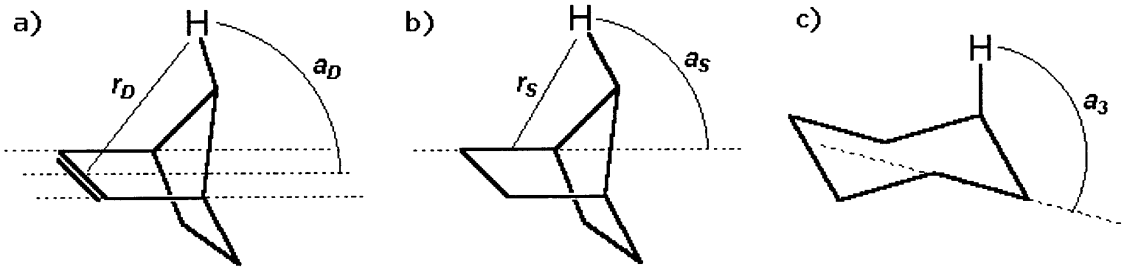


Figure 3. Special distance measures for the characterization of proton environments. (a) distance r_D and radian angle a_D to double bonds; (b) distance r_S and radian angle a_S to single bonds; (c) dihedral radian angle a_3 to the third bond from the hydrogen atom.

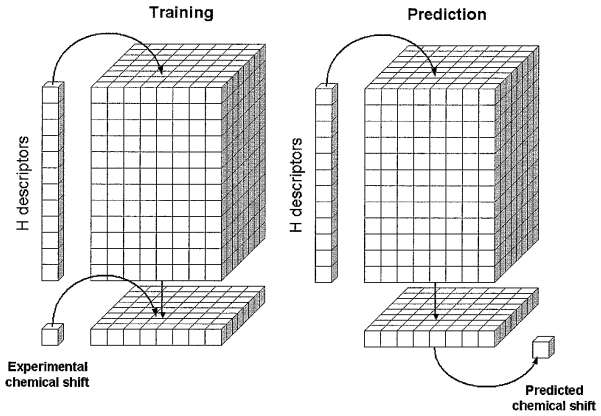


Figure 4. Schematic representation of a CPG NN. The CPG NN is trained with proton descriptors and the corresponding observed ^{1}H NMR chemical shifts. After training, the NN is able to predict the chemical shift on input of proton descriptors.

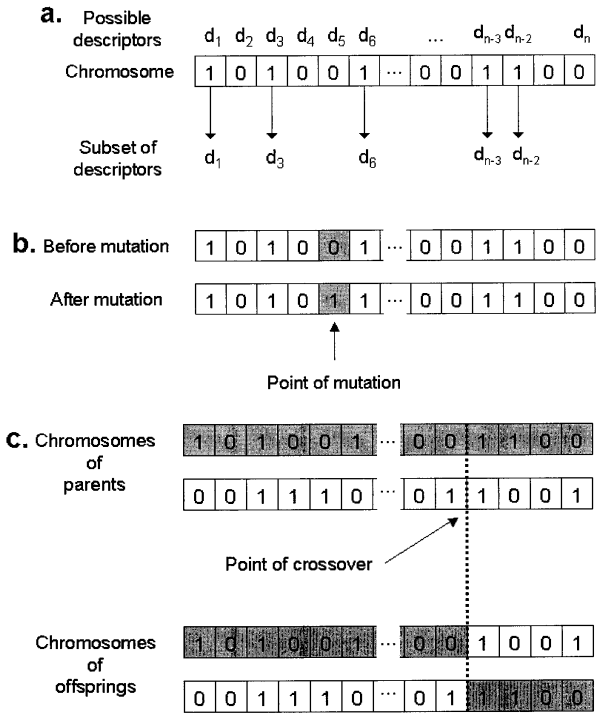


Figure 5. (a) Specification of a subset of descriptors from a pool of possible descriptors d_1−d_n by a chromosome. (b) Mutation of one gene. (c) Crossover between chromosomes of parents and chromosomes of the resulting offsprings.

considered. The collection was restricted to CH_n protons and to compounds containing elements C, H, N, O, S, F, Cl, Br, or I. Hydrogen atoms bonded to heteroatoms were not included as their chemical shifts are strongly influenced by experimental conditions (such as concentration of the sample). The data sets were designed in order to cover as many situations of protons belonging to organic structures as possible.

Four different classes of protons were defined (aromatic, nonaromatic π, rigid aliphatic, nonrigid aliphatic), and protons belonging to each of them were treated separately. This procedure allowed the use of more specific descriptors for each class.

Protons were classified as (a) aromatic when they are bonded to an aromatic system, (b) nonaromatic π when they are bonded to a nonaromatic π system, (c) rigid aliphatic when a nonrotatable bond is identified in the second sphere of bonds centered on the proton, and (d) nonrigid aliphatic when not included in previous classes. A bond was defined as nonrotatable if it belongs to a ring, to a π system, or to an amide functional group. An example of classification of protons in a molecule is given in Figure 1.

The training set was divided accordingly into four training sets (one for each class), and the same procedure was applied to the prediction set. The number of cases in the training and prediction sets were respectively 145 and 39 for aromatic protons, 117 and 47 for π protons, 237 and 87 for rigid aliphatic protons, and 245 and 86 for nonrigid aliphatic protons.

Representation of Protons. To be submitted to a neural network, each proton was represented by a fixed number of descriptors. Physicochemical, geometric, and topological descriptors were calculated.

Physicochemical descriptors were based on empirical values calculated by the software package PETRA[19] (version 3.0) comprising a variety of previously published methods[19a] for the protons and for their neighborhood. Examples of physicochemical descriptors used in this study are as follows: partial atomic charge of the proton, effective polarizability of the proton, average of partial atomic charges of atoms in the second sphere, maximum partial atomic charges of atoms in the second sphere, minimum effective polarizability of atoms in the second sphere, and average of σ electronegativities of atoms in the second sphere.

Geometric descriptors were based on the 3D molecular structure generated by the CORINA software package.[20] The

H. *J. Nat. Prod.* **1996**, *59*, 1192−1195. (m) Chen, H. C.; Zhu, Y.; Shen, X.-M.; Jia, Z.-J. *J. Nat. Prod.* **1996**, *59*, 1117−1120. (n) Kuo, Y.-H.; Yeh, M.-H. *Phytochemistry* **1998**, *49*, 2453−2455. (o) SDBS database, available at http://www.aist.go.jp/RIODB/SDBS/menu-e.html (accessed between January 2000 and January 2001).

(19) (a) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, Germany, 1988; pp 119−138. (b) http://www2.chemie.uni-erlangen.de/software/petra.

global 3D environment of a proton belonging to a rigid substructure was represented by a proton RDF of the form

$$g_H(r) = \sum_i p_i e^{-B(r-r_i)^2} \quad (1)$$

where i denotes an atom up to four nonrotatable bonds away from the proton under consideration, p_i is the partial atomic charge of atom i, B is a smoothing parameter, r_i is the 3D distance between the proton and the atom i, and r is the running variable. In the plot of $g_H(r)$ against r, each 3D distance contributes to a peak, and the contribution is proportional to p_i. An example is shown in Figure 2. Values of $g_H(r)$ at fixed points are used as descriptors of the proton.

Other more specific geometric features were represented using functions analogous to eq 1 but encoding other properties. The influence of the electronic current in double bonds was incorporated by $g_D(r)$, a function similar to eq 1, where i denotes now a double bond up to the seventh sphere of nonrotatable bonds centered on the proton, while the geometric variables r_D and a_D are illustrated in Figure 3a.

$$g_D(r) = \sum_i \frac{1}{r_{D,i}^2} e^{-B(r-a_{D,i})^2} \quad (2)$$

Shielding and unshielding by single bonds was encoded using $g_S(r)$, where i denotes a single bond up to the seventh sphere of nonrotatable bonds centered on the proton, the geometric variables r_S and a_S are respectively the distance and the radian angle represented in Figure 3b.

$$g_S(r) = \sum_i \frac{1}{r_{S,i}^2} e^{-B(r-a_{S,i})^2} \quad (3)$$

To account for axial and equatorial positions of protons bonded to cyclohexane-like rings, $g_3(r)$ was used, where i denotes an atom three nonrotatable bonds away from the proton and belonging to a six-membered ring and a_3 is a dihedral angle in radian units (Figure 3c).

$$g_3(r) = \sum_i e^{-B(r-a_{3,i})^2} \quad (4)$$

The minimum and the maximum bond angles centered on the atom adjacent to the proton were also used as geometric descriptors. For aromatic and nonrigid aliphatic protons, these were the only two geometric descriptors used. In addition to these, for nonaromatic π protons, the RDF function $g_H(r)$ and the number of non-hydrogen atoms at cis and trans positions were used as geometric descriptors, as we were mostly interested in predicting the influence of cis/trans isomerism. For rigid aliphatic protons, all the geometric descriptors were calculated.

Table 1. Prediction of ^{1}H NMR Chemical Shifts by CPG Neural Networks of Different Sizes Using All Calculated Descriptors[a]

type of protons	size of CPG NN	mean absolute error (ppm) in training set	mean absolute error (ppm) in prediction set	prediction of stereochemical effect[b] (prediction set)
aromatic	12 × 12	0.09	0.31	-
	13 × 13	0.09	0.31	
	14 × 14	**0.08**	**0.23**	
	15 × 15	0.06	0.25	
	16 × 16	0.06	0.29	
	17 × 17	0.05	0.27	
nonrigid aliphatic	15 × 15	0.07	0.28	-
	16 × 16	0.07	0.29	
	17 × 17	0.05	0.24	
	18 × 18	0.05	0.23	
	19 × 19	0.04	0.25	
	20 × 20	0.04	0.27	
	21 × 21	**0.03**	**0.21**	
	22 × 22	0.03	0.23	
π-bonded	11 × 11	0.10	0.26	6/7
	12 × 12	0.07	0.26	6/7
	13 × 13	0.06	0.26	5/7
	14 × 14	**0.05**	**0.22**	**5/7**
	15 × 15	0.04	0.31	6/7
rigid aliphatic	15 × 15	0.16	0.39	16/28
	16 × 16	0.18	0.35	18/28
	17 × 17	0.14	0.35	20/28
	18 × 18	0.13	0.34	17/28
	19 × 19	**0.11**	**0.34**	**21/28**
	20 × 20	0.10	0.35	18/28
	21 × 21	0.09	0.37	21/28
	22 × 22	0.10	0.33	18/28

[a] Optimum sizes displayed in boldface type. [b] Number of correct predictions/number of total cases. Each case is a CH_2 group in which the two protons are distinguishable. Correct prediction is assumed if the order of the two predicted chemical shifts is correct or if the two protons have the same chemical shift and a difference of less than 0.1 ppm was predicted.

Topological descriptors were based on the analysis of the connection table, and the properties were calculated by the methods contained in PETRA. Some examples of topological descriptors are as follows: number of carbon atoms in the second sphere centered on the proton, number of oxygen atoms in the third sphere, and number of atoms in the second sphere that belong to an aromatic system. A topological $g_H(r)$ function was also used, where i denotes any atom up to the fifth sphere centered on the proton, p_i is the partial atomic charge of atom i, and r_i is now the sum of bond lengths on the shortest possible path between the proton and the atom i.

Some topological and physicochemical descriptors were not used for some classes if they had the same values for almost all the protons of that class. All in all, 92 descriptors were used for aromatic protons, 119 for nonrigid aliphatic protons, 101 for π protons, and 174 for rigid aliphatic protons.

Counterpropagation Neural Networks. To model the relationship between the proton descriptors and the corresponding chemical shifts, CPG NN[14c] were used. CPG networks were chosen because they are especially useful for the modeling of complex and nonlinear relationships.

The input data for a CPG network are stored in a two-dimensional grid of neurons, each containing as many elements

(20) (a) Sadowski, J.; Gasteiger, J. *Chem. Rev.* **1993**, *93*, 2567−2581. (b) Gasteiger, J.; Rudolph, C.; Sadowski, J. *Tetrahedron Comput. Methodol.* **1992**, *3*, 537−547. (c) Sadowski, J.; Rudolph, C.; Gasteiger, J. *Anal. Chim. Acta* **1992**, *265*, 233−241. (d) Sadowski, J.; Gasteiger, J.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

Table 2. Prediction of ^{1}H NMR Chemical Shifts by CPG Neural Networks (Using Descriptors Selected by Genetic Algorithms)

			training set		prediction set		
type of proton	size of CPG NN	no. of descriptors	no. of cases	mean absolute error (ppm)	no. of cases	mean absolute error (ppm)	prediction of stereochemical effect[a]
aromatic	14 × 14	20	145	0.07	39	0.21	
nonrigid aliphatic	21 × 21	17	245	0.04	86	0.19	
π-bonded	14 × 14	19	117	0.03	47	0.24	4/7
rigid aliphatic	19 × 19	42	237	0.09	87	0.37	17/28

[a] See footnote *b* in Table 1.

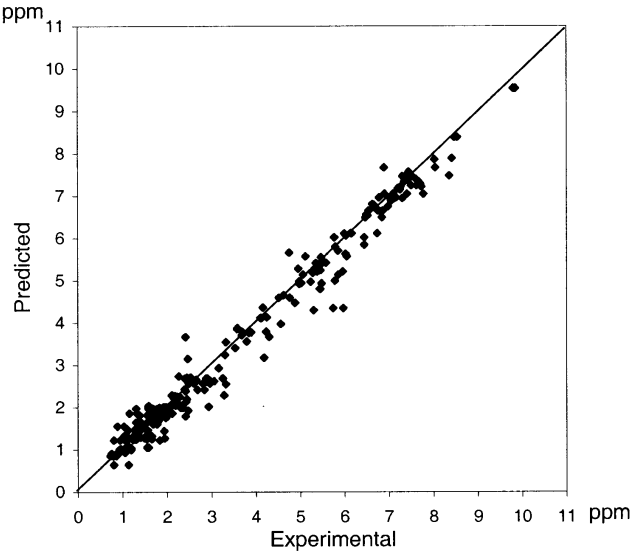


Figure 6. Observed chemical shifts of protons plotted against those predicted by the neural network.

(weights) as there are input variables. In the investigations described in this paper, the input variables are proton descriptors. In Figure 4, this part of the CPG network is represented by the upper block, which is basically a Kohonen[21] network. The output data (in this case the chemical shifts) are stored in a second layer that acts as a lookup table.

Before the training of a CPG network starts, random weights are generated. During the training, each individual object (proton) is mapped into that neuron of the Kohonen layer (central neuron or winning neuron) that contains the most similar weights compared to the input data (descriptors). The weights of both input and output layers are then adjusted to make them even more similar to the presented data. The extent of adjustment depends on the topological distance to the central neuron. The network is trained iteratively, i.e., all the objects of the training set are presented several times, and the weights are corrected, until the network stabilizes. Note that chemical shifts are not used in determining the winning neuron. Thus, it is not of influence on the final mapping; we have an unsupervised learning technique.

After the training, the CPG NN is able to predict the chemical shift on input of a proton represented by its descriptors. The winning neuron is chosen and the correspondent value in the output layer is used for prediction (Figure 4).

(21) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.

Selection of Variables Using Genetic Algorithms. For the selection of variables, evolution of a population was simulated. Each individual of the population represented a subset of descriptors and was defined by a chromosome of binary values. The details of the method are described below.

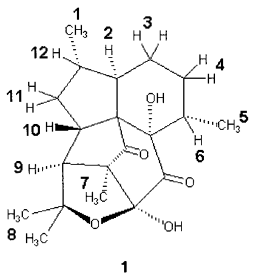
The chromosome has as many genes as there are possible descriptors (92 for the aromatic group, 119 for nonrigid aliphatic, 174 for rigid aliphatic, and 101 for nonaromatic π protons), each gene corresponding to one descriptor. One gene takes a value of 1 if the corresponding descriptor is included in the subset, and it takes a value of 0 if the descriptor does not belong to the subset represented by the individual (Figure 5a).

At the beginning of the evolution, the chromosomes are randomly generated. In order that the number of genes with value 1 is kept relatively low (small subsets of descriptors), the probability of generating 0 for a gene was set (randomly) between 2.3 and 5.7 times higher than that of generating 1 for a gene.

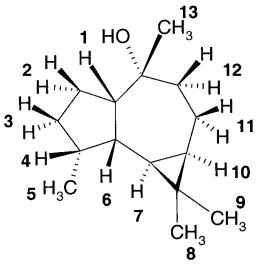
A population size of typically 250 individuals was chosen, and evolution was allowed over typically 50 generations. In each generation, half of the individuals mate (the fittest individuals), and the other half die. Each of the surviving individuals mates with another (randomly chosen) surviving individual, and two new offsprings are generated. The chromosomes of the offsprings result from crossover of their parents' chromosomes, followed by mutation (Figure 5). The population in the next generation is made by the new offsprings and their parents.

Crossover occurs at a randomly chosen single point. Mutation is allowed to occur at every gene of the new offspring with a random probability. The probability of mutation 0 → 1 is set for each individuum (randomly) between 0 and 5%. The probability of mutation 1 → 0 is set for each individuum (randomly) between 4 and 6 times higher than the probability of mutation 0 → 1.

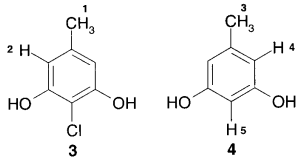
The evaluation (scoring) of each chromosome is made by a CPG neural network that uses the subset of descriptors encoded in the chromosome for predicting chemical shifts. This NN works like those described above. Each proton is fed to the NN as a set of descriptors (input) obtaining a chemical shift as output. The NN is trained with a subset of the training set (reduced training set) and applied to the cross-validation set (the remaining subset of the training set). The score of one chromosome (fitness function) is the root mean square of errors for the chemical shifts obtained with the cross-validation set. Chromosomes with lower



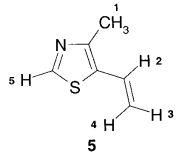
No.	Observed	NN	ACD I-Lab	TopNMR
1	0.99	0.97	0.978	1.05
2	2.03	2.39	1.853	1.62
3α	1.81	2.05	2.239*	1.39
3β	1.53	1.97	1.366*	1.39
4α	1.26	1.65	1.782*	1.39
4β	2.04	1.97	1.857*	1.39
5	0.62	0.97	1.068	1.05
6	2.12	2.39	2.051	1.90
7	1.42	1.27	1.353	1.25
8	1.40	1.27	1.566	1.25
9	1.99	2.01	2.252	2.01
10	3.17	2.39	2.138	1.98
11α	0.65	1.45	1.605*	1.42
11β	1.94	1.97	2.008*	1.42
12	1.68	2.27	2.524	1.66
Mean absolute error		0.303	0.341	0.346



No.	Observed	NN	ACD I-Lab	TopNMR
1	2.09	2.01	2.006	1.55
2α	1.90	1.76	2.118*	1.47
2β	1.69	1.65	1.918*	1.47
3α	1.30	1.58	1.384*	1.47
3β	1.70	1.78	1.645*	1.47
4	1.99	2.27	2.243	1.66
5	0.94	0.91	0.92	1.05
6	1.78	1.71	1.597	1.47
7	0.33	1.73	0.786	0.17
8	0.98	0.86	0.995	1.11
9	1.04	0.86	1.025	1.11
10	0.72	1.73	0.757	0.18
11α	1.83	1.76	1.585*	1.25
11β	1.21	1.76	1.42*	1.25
12α	1.86	1.69	1.483*	1.39
12β	1.69	1.58	1.226*	1.39
13	1.14	1.27	1.19	1.30
Mean absolute error		0.279	0.176	0.282



No.	Observed	NN	ACD I-Lab	TopNMR
1	2.24	2.25	2.367	2.34
2	6.43	6.43	6.348	6.03
3	2.23	2.25	2.367	2.34
4	6.24	6.81	6.159	6.09
5	6.17	6.45	5.994	6.01
Mean absolute error		0.176	0.126	0.184



No.	Observed	NN	ACD I-Lab	TopNMR
1	2.44	2.20	2.29	2.47
2	6.77	6.65	6.657	6.63
3	5.23	4.97	5.254	5.17
4	5.45	5.24	5.528	5.61
5	8.53	8.42	8.602	8.7
Mean absolute error		0.188	0.087	0.112

Figure 7. Prediction of ^{1}H NMR chemical shifts by CPG NN in comparison to results obtained from the ACD I-Lab and TopNMR program packages. Asterisk values: The ACD/I-Lab interface did not allow unambiguous assignment of the chemical shifts to α- and β-protons. In these tables, and for the calculation of errors, it was assumed that the order of the chemical shifts was correctly predicted.

scores are considered to be fitter than those with higher scores and are selected for mating.

For each of the four groups of protons, the reduced training set and the cross-validation set were chosen by a Kohonen neural network. The Kohonen NN was trained with all the protons of the training set, characterized by all the possible descriptors. After the training, half of the protons mapped into a neuron was assigned to the reduced training set and the other half was assigned to the cross-validation set. This was done for each neuron, and when there was an odd number of protons in a neuron, the last proton was randomly assigned.

Computational Details. The Cartesian coordinates of the atoms in a molecule were calculated from the connection tables of the molecules by the 3D structure generator CORINA.[20] The physicochemical atomic properties were calculated using fast empirical methods implemented in the program package PETRA 3.0.[19] For the calculation of $g_H(r)$, $g_D(r)$, $g_S(r)$, and $g_3(r)$ the smoothing parameter B was set to 20, 1.15, 1.15, and 2.86. Each descriptor (except RDF descriptors) was linearly scaled between 0 and 1 for the cases of the training set, and the same scaling factors were applied to the prediction set. RDF functions were vector sum normalized. $g_H(r)$ was calculated at 15 evenly distributed points between 1.4 and 4 Å. $g_D(r)$ and $g_S(r)$ were calculated at seven evenly distributed points between 0 and $\pi/2$. $g_3(r)$ was calculated at 13 evenly distributed points between 0 and π. The classification of protons and the calculation of descriptors were performed by computer programs especially developed for this task. The programs were written using the C programming language and were compiled for WindowsNT and SGI IRIX 6.3 platforms.

The CPG networks used in this investigation were simulated with the CPG network simulator KMAP (version 4.0).[22] Network topologies of toroidal shape were chosen. Training of the CPG network was performed by using a linear decreasing triangular scaling function used with an initial learning span of 8 and an initial learning rate of 0.7. The weights were initialized with normally distributed pseudorandom numbers that were calculated using the mean and standard deviation of the input data. For the selection of the central neuron, the minimum Euclidean distance between the input vector and neuron weights was used. Training was performed until the learning span was reduced to 0 and the learning rate was reduced to 0.1, in unsupervised manner; i.e., the values in the output layer were adapted but not used for the selection of the central neuron. After every n iterations ($n = 20$ in the GA experiments and $n =$ number of training objects in the other experiments), the sum of errors in the input layer was calculated and compared with the previous error. When it was lower, the learning rate was multiplied by 0.95.

The whole process to predict the ^{1}H NMR shifts of 30 protons in a molecule with 56 atoms, starting from an MDL Molfile and using the best models, took 2 s on an SGI Origin 200 workstation with one 180-MHz IP27 processor and running IRIX 6.3.

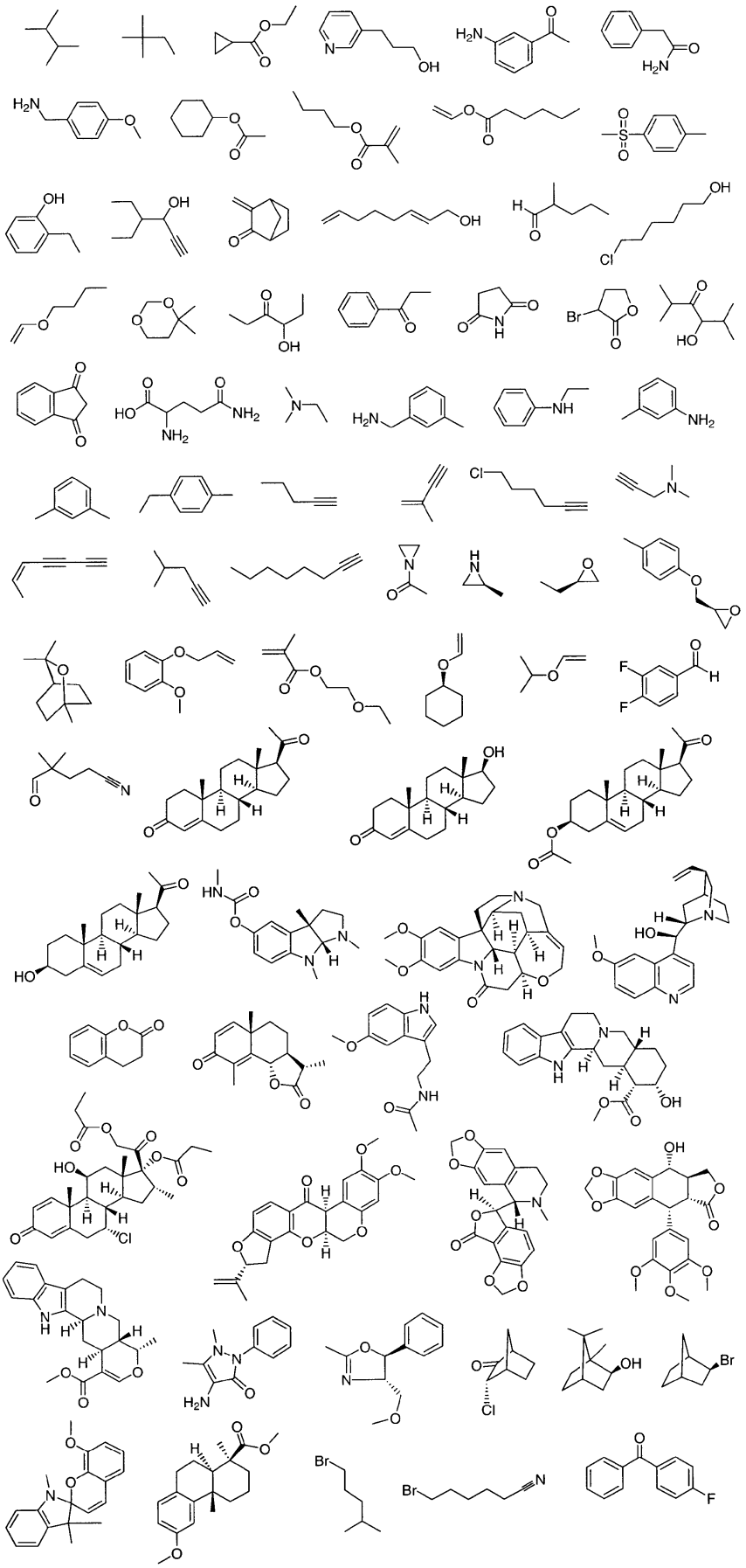
Evaluation of a chromosome coding the expression of 20 descriptors, using a 9 × 9 CPG NN, a training set with 77 objects, and a validation set with 39 objects took 0.7 s on the same hardware.

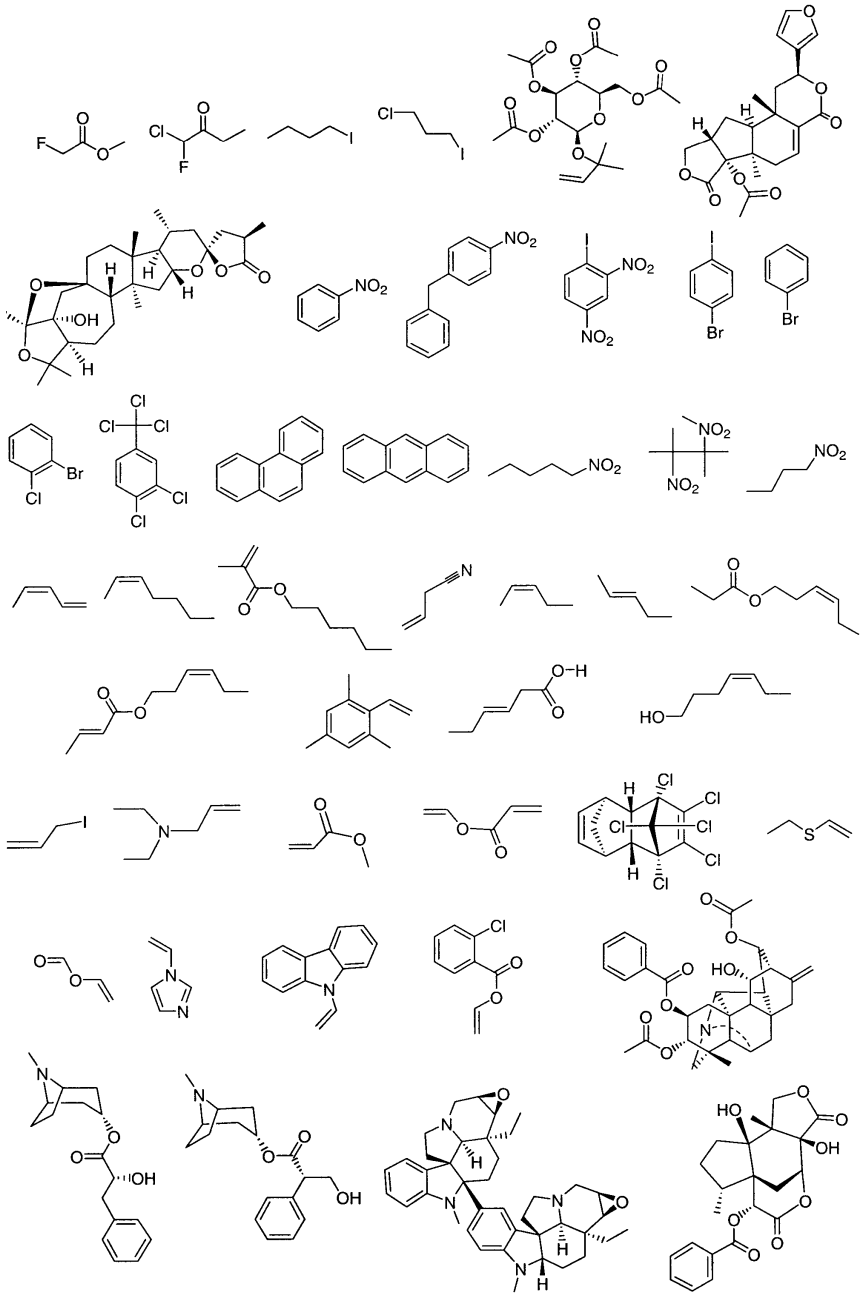
RESULTS AND DISCUSSION

CPG NN Using All Calculated Descriptors. Counterpropagation neural networks were trained to make predictions of ^{1}H

(22) Teckentrup, A. Ph.D. Thesis, University of Erlangen-Nürnberg, Erlangen, Germany, 2000.

Chart 1. Structures from Which the Training Set Was Extracted





NMR chemical shifts for protons of the type CH_n. Protons belonging to four different classes of substructures were treated separately: aromatic, π nonaromatic, rigid aliphatic, and nonrigid aliphatic (see Methodology).

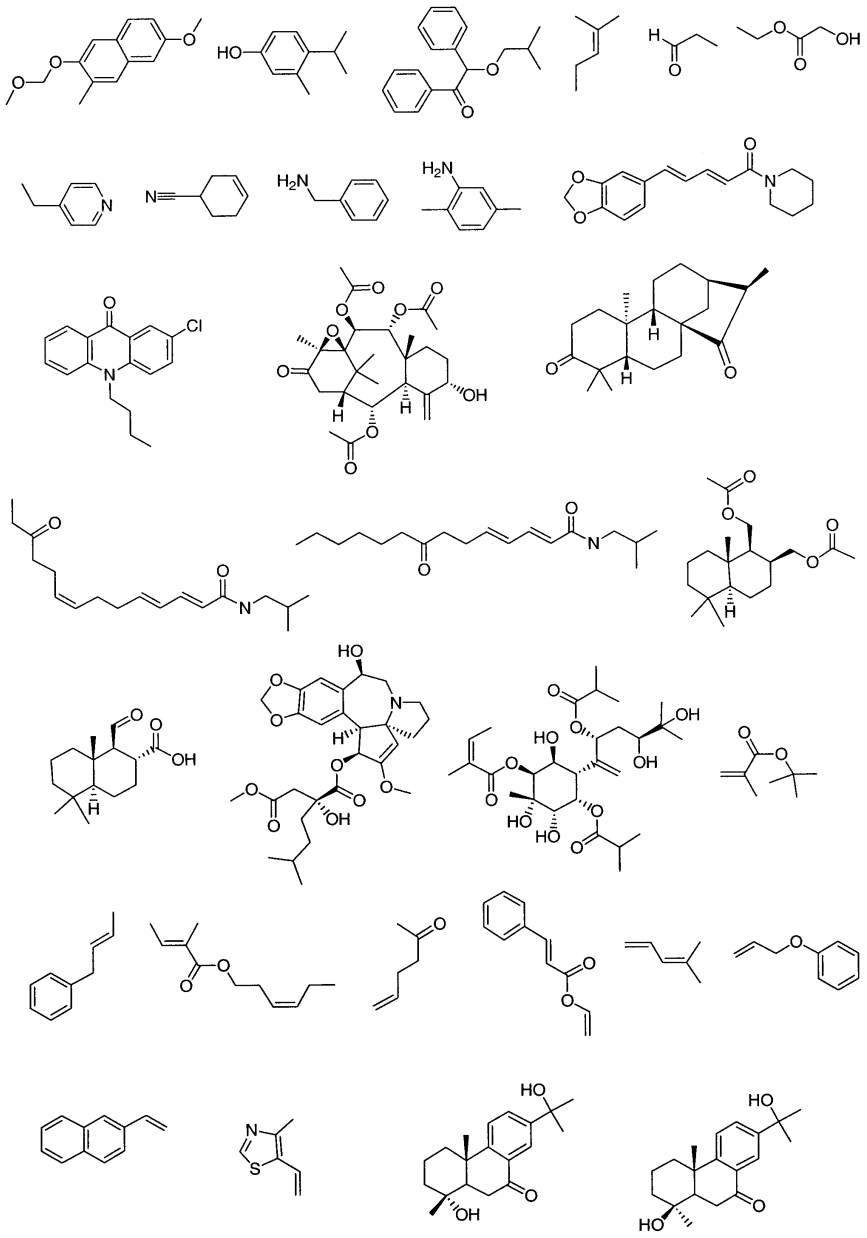
All descriptors generated for each type of proton were used, and different sizes of neural networks were tested with the number of neurons varying approximately between the number of cases in the training set and 1.5 times this value. The networks were trained with the training sets and tested with the prediction sets (Table 1).

Excellent predictions were obtained for protons of aromatic, π, and nonrigid aliphatic classes. Although higher, the errors for the rigid aliphatic type are also good considering the complexity of the structures used and the difficulty in accounting for 3D effects.

A measure of how well stereochemical effects are accounted for cannot easily be evaluated. One possible way is to analyze the predictions for CH_2 protons in cases where the two protons are distinguishable (for stereochemical reasons). In our approach, this was possible for nonaromatic π and for rigid aliphatic protons. We followed this approach by assuming a true prediction when the network was able to correctly decide which of the two CH_2 protons has a higher chemical shift. This is a rather difficult task for the rigid aliphatic class, as the average difference in chemical shifts for the distinguishable CH_2 protons in the prediction set is 0.31 ppm. Despite that fact, true predictions for more than 50% of the cases were obtained with all the networks tested (Table 1). For the rigid aliphatic class, the best result was 75%. These results indicate that the geometric descriptors used can represent stereochemical and 3D effects in a meaningful way.

The optimum size for the neural network in each of the four classes (displayed in boldface type in Table 1) was chosen on the basis of mean absolute error and prediction of stereochemical effects for the prediction set. The optimum sizes were used in

Chart 2. Structures from Which the Prediction Set Was Extracted



further experiments described below.

CPG NN Using Variables Selected by GA. To obtain more compact, robust, and accurate models, CPG neural networks were trained with subsets of descriptors, instead of all generated descriptors for each class. The selection of the subsets was done by genetic algorithms as described in the Methodology section. The selected descriptors were used to train and test CPG NN with the same size, the same training set, the same prediction set, and the same training parameters as the best networks selected in Table 1. The results are displayed in Table 2.

Comparing with the predictions using all the descriptors, the results for aromatic and for nonrigid aliphatic protons were slightly improved (0.02 ppm in mean absolute error) using selected descriptors. But the new models are quite convenient since much less descriptors are required (20 instead of 92 descriptors for aromatic protons, and 17 instead of 119 descriptors for nonrigid aliphatic protons).

A different situation occurred with protons belonging to π and rigid aliphatic substructures. For these groups, the selected descriptors yielded slightly poorer results compared to the full set of descriptors, both in terms of mean absolute error (0.24 vs 0.22 ppm for π; 0.37 vs 0.34 ppm for rigid aliphatic) and prediction of stereochemical effects (4/7 vs 5/7 for π; 17/28 vs 21/28 for rigid aliphatic), see Tables 1 and 2.

Application of the Best Models. For further applications, it was decided to choose the best models for each of the four classes of protons: for aromatic protons, a 14 × 14 neural network and the subset of 20 descriptors selected by GA; for nonrigid aliphatic protons, a 21 × 21 neural network and the subset of 17 descriptors selected by GA; for π protons, a 14 × 14 neural network and all the descriptors defined for the class; and for rigid aliphatic protons, a 19 × 19 neural network using all the descriptors defined for the class.

Table 3. List of Descriptors Used in the Work

type of descriptor	properties used
properties of the proton	partial atomic charge, effective polarizability
properties of the atom bonded to the proton	partial atomic charge, π atomic charge, σ atomic charge, π electronegativity, σ electronegativity, effective polarizability
properties of the covalent bond involving the proton	delocalization stabilization of a positive charge, delocalization stabilization of a negative charge, resonance stabilization, difference in σ electronegativity between the two atoms of the bond, difference in π electronegativity between the two atoms of the bond, difference in π atomic charge between the two atoms of the bond, difference in σ atomic charge between the two atoms of the bond, difference in partial atomic charge between the two atoms of the bond, difference in effective polarizability between the two atoms of the bond
minimum, maximum, and average of properties of atoms two bonds away from the proton	partial atomic charge, π atomic charge, σ atomic charge, π electronegativity, σ electronegativity, lone-pair electronegativity, effective polarizability, number of free electrons
count (or percentage) of atoms two bonds away from the proton with specific identity or properties	carbon atom, hydrogen atom, oxygen atom, nitrogen atom, aromatic atom, π atom, atom belonging to a cycle
minimum, maximum, and average properties of covalent bonds in the second sphere of bonds (from the proton)	delocalization stabilization of a positive charge, delocalization stabilization of a negative charge, delocalization stabilization, resonance stabilization, difference in σ electronegativity between the two atoms of the bond, difference in π electronegativity between the two atoms of the bond, difference in π atomic charge between the two atoms of the bond, difference in σ atomic charge between the two atoms of the bond, difference in partial atomic charge between the two atoms of the bond, difference in effective polarizability between the two atoms of the bond
count of atoms three bonds away from the proton with specific identity or properties	carbon atom, hydrogen atom, oxygen atom, nitrogen atom, aromatic atom, π atom, nonaromatic π atom, nonaromatic atom
count of atoms four or five bonds away from the proton (within an aromatic system)with specific identity or properties	aromatic atom, nonaromatic atom
RDF functions	$g_H(r)$ with 3D distances, $g_H(r)$ with topological distances, $g_D(r)$, $g_S(r)$ for single bonds with H atom, $g_S(r)$ for single bonds with no H atom, $g_3(r)$
other	shielding or deshielding by electronic current of aromatic system (only in rigid substructures); minimum and maximum angle of bonds involving the proton.

Using these models, the global mean absolute error for the prediction set is 0.25 ppm with standard deviation of 0.25 ppm. A mean absolute error of 0.19 ppm was obtained for 90% of the cases. A plot of the predicted versus the observed chemical shifts is displayed in Figure 6.

Application of the chosen trained networks is illustrated with molecules **1**−**5**[18o,23−25] that were not present in the training set. In Figure 7, the predictions are shown and compared with the results of commercial software (ACD I-Lab[5] and TopNMR[9]). These structures were chosen to represent different types of protons. Molecules **1**−**4** do not belong to the prediction set and are new to all three methods, since they were only recently described in the literature. Of particular interest is elisapterosin A (**1**), which possesses a previously unprecedented cagelike framework.[23]

Although these five molecules are obviously too few for a reliable and comprehensive comparison of the methods, the results can give some hints about the potential of the new method.

For structure **1**, the NN method outperforms the other two in terms of mean absolute error. Furthermore, for all the diastereotopic pairs of protons of the type CH_2 belonging to rigid substructures, the order of chemical shifts was correctly predicted. This is a unique feature of the NN method, as TopNMR does not distinguish diastereotopic protons. This ability of the neural network is due to the use of geometric descriptors for the representation of protons. ACD I-Lab distinguishes between diastereotopic protons, but for a structure of such complexity, unambiguous assignments of the predicted values could not be obtained, probably due to interface limitations.

Stereochemical effects in rigid substructures are also reasonably well predicted for structure **2**.[24] However, a poor prediction was obtained for the protons attached to cyclopropane in this compound. It can be explained by the absence of similar structures in the training set—only one cyclopropane derivative was used for training, and this has a carbonyl group adjacent to the small ring. If this case is excluded for the calculation of the mean absolute errors, the global performance for molecules **1**−**5** is approximately the same for the neural network method and for ACD I-Lab (0.22 ppm), while TopNMR gave slightly worse predictions (0.27 ppm). This example illustrates the importance of the training set. It should be noted at this point that ACD I-Lab used a database of 800 000 ^{1}H NMR chemical shifts[5] while we trained the neural networks with a data set of only 744 cases.

An example of remarkable behavior of the NN is given by the hydrogen atom bonded to the heterocycle of structure **5**.[18o] A very good prediction (8.53 vs 8.42 ppm) was obtained with the NN, although no heterocycle containing sulfur was used for training. Such predictions in new situations are made possible by the use of physicochemical and topological descriptors that generalized atom types to inherent physicochemical properties.

As a whole, the results for structures **1**−**5** indicate that the neural network method, *even at the present stage of development*, can make predictions of at least the same quality as those of commercial packages. The predictions obtained for **1** and **2** suggest that, in rigid structures where 3D effects are strong, the new approach has the potential to outperform the available methods.

(23) Rodriguez, A. D.; Ramirez, C.; Rodriguez, I. I.; Barnes, C. L. *J. Org. Chem.* **2000**, *65*, 1390−1398.
(24) Kaplan, M. A. C.; Pugialli, H. R. L.; Lopes, D.; Gottlieb, H. E. *Phytochemistry* **2000**, *55*, 749−753.
(25) Monde, K.; Satoh, H.; Nakamura, M.; Tamura, M.; Takasugi, M. *J. Nat. Prod.* **1998**, *61*, 913−921.

Table 4. List of Descriptors Selected by GA for Aromatic Protons

π atomic charge of the adjacent atom to the proton
resonance stabilization of the covalent bond involving the proton
difference in π atomic charge between the proton and the adjacent atom
average of σ electronegativity of atoms two bonds away
average of the difference in σ electronegativities in bonds of the second sphere
average of the difference in π atomic charges in bonds of the second sphere
maximum π atomic charge of atoms two bonds away
minimum difference in σ electronegativities in bonds of the second sphere
minimum difference in π electronegativities in bonds of the second sphere
minimum difference in π atomic charges in bonds of the second sphere
number of carbon atoms two bonds away
number of oxygen atoms three bonds away
number of nonaromatic atoms three bonds away
number of π nonaromatic atoms three bonds away
number of aromatic atoms four bonds away (within aromatic system)
topological $g_H(r)$ at 1.6, 3.0, 4.0, 4.2, and 5.6 Å

Table 5. List of Descriptors Selected by GA for Nonrigid Aliphatic Protons

difference in σ electronegativity between the proton and the adjacent atom
average of partial atomic charge of atoms two bonds away
average of the difference in π electronegativities in bonds of the second sphere
maximum π atomic charge of atoms two bonds away
maximum σ electronegativity of atoms two bonds away
maximum lone-pairs electronegativity in atoms two bonds away
maximum resonance stabilization of a negative charge by bonds of the second sphere
maximum difference in σ electronegativities in bonds of the second sphere
minimum partial atomic charge of atoms two bonds away
minimum σ electronegativity of atoms two bonds away
minimum delocalization stabilization by bonds of the second sphere
minimum difference in σ atomic charges in bonds of the second sphere
number of nitrogen atoms two bonds away
number of pi atoms two bonds away
number of hydrogen atoms three bonds away
topological $g_H(r)$ at 1.2 and 4.8 Å

A particularly useful feature of the neural network approach is that the system can be easily retrained for specific types of compounds. Assigned spectra of related structures can be added to the training set, and the training can be repeated by using the same descriptors or by selecting other ones from the pool of available descriptors. Improved results can be expected for similar compounds.

CONCLUSION

A new system based on CPG NN was developed for fast estimation of NMR chemical shifts of CH_n protons in organic compounds. The use of physicochemical, topological, and geometric descriptors was revealed to be a quite successful strategy. The system achieved mean absolute errors of prediction between 0.2 and 0.3 ppm for the prediction set and for independent structures of four new natural products recently discovered. In comparison with two of the most well-known commercial packages, the neural networks gave errors approximately in the same range, with the advantage of distinguishing and assigning some diastereotopic protons. Particularly better results were obtained for rigid structures, for which geometric descriptors have been calculated.

The geometric descriptors were based on the calculated 3D molecular structure of a single conformation and allowed the storage of information concerning the influence of the proton 3D environment on its ^{1}H NMR chemical shift.

Most of the less accurate predictions could be related to the absence of similar situations in the training set. This fact suggests that a customized system can be trained for a specific type of compounds. Retraining the networks with a specific data set should improve the prediction quality for related structures.

Globally, the performance of the method, at the current stage of development, is remarkable if one considers the relatively small data set on which it was based. Work is in progress using a much larger database for training, and improvements can be expected.

APPENDIX A

Chart 1 shows the structures from which the training set was extracted.

APPENDIX B

Chart 2 shows the structures from which the prediction set was extracted.

APPENDIX C

This appendix contains the lists of descriptors used in this work (Table 3), selected by GA for aromatic protons (Table 4), and selected by GA for nonrigid aliphatic protons (Table 5).

ACKNOWLEDGMENT

J.A.-d.-S. acknowledges Fundação para a Ciência e a Tecnologia (Lisboa, Portugal) for a postdoctoral fellowship. J.G. thanks the Bundesministerium für Bildung und Forschung (bmb+f; grant 0311680) and the Fonds der Chemischen Industrie for financial support.

Received for review July 2, 2001. Accepted October 5, 2001.

AC010737M