Rapid Access to Infrared Reference Spectra of Arbitrary Organic Compounds: Scope and Limitations of an Approach to the Simulation of Infrared Spectra by Neural Networks

Paul Selzer,[a] Johann Gasteiger,*[a] Henrik Thomas,[b] and Reiner Salzer[b]

Abstract: Substance identification by infrared spectroscopy is performed by comparison of the experimental spectrum with a reference spectrum from a printed compilation or a database. If the analyzed compound can not be found in a database the corresponding reference spectrum has to be simulated. In order to achieve this, several reasonable candidates of structures for the compound at hand have to be conceived and for all these, infrared spectra have to be developed. The simulated spectrum that is most similar to the experimental suggests the correct structure. A rapid spectrum prediction method based on neural networks has been developed that supplies reference spectra for any organic compound. The scope and limitations of this method will be discussed on a test set of 16 compounds representing a broad range of organic chemistry.

Keywords: analytical chemistry · infrared spectroscopy · neural networks · spectrum simulation

Introduction

Infrared spectra can play an important role in the identification of organic compounds. Infrared spectroscopy is a nondestructive method requiring only small amounts of a sample. Furthermore, an infrared spectrum has a high information content which is quite specific for a particular compound. Not without reason a specific range of an infrared spectrum is called fingerprint region to stress its significance for the identification of a chemical compound.

The identification of a compound asks for the comparison of its infrared spectrum with the reference spectrum of this compound as taken from a printed compilation or from a database. There the problem arises: In comparison with the number of known compounds (more than 16 000 000), the number of infrared spectra in databases is rather small (100 000 in the largest database). In order to overcome this discrepancy, infrared spectra for compounds not included in a database have to be simulated.[1–3] Infrared spectra can be calculated by quantum mechanical methods. However, in order to achieve a good agreement with experiments ab initio calculations with quite large basis sets or a density functional theory approach are required asking for substantial computational time. We wanted to provide a more rapid approach to infrared spectra in order to be able to generate large sets of reference spectra within a short time. The relationships between an infrared spectrum and chemical structure have then to be learnt from data. In this endeavor, we wanted to provide high quality infrared spectra covering the entire spectral range, including the highly significant fingerprint region which is important for structure identification. This region shows deformation and skeletal vibrations and strong couplings between these vibrations. This requirement precluded the use of a fragment-based approach which can only provide a few of the bands observed in an infrared spectrum, predominantly valence vibrations.

Infrared spectroscopy monitors the movements of the atoms of a molecule in 3D space. Thus, any approach to learning the relationships between infrared spectra and chemical structure should first of all start from a representation of the 3D structure. We have recently developed a novel representation of the 3D structure, the 3D-MoRSE code (3D-molecule representation of structures based on electron diffraction).[4, 5] Based on this work we then introduced a coding of the 3D structure of molecules by atom radial distribution functions because we have found a way to interpret this representation to obtain again a 3D structure.[6] We will demonstrate here the merits of this structure coding method for the simulation of infrared spectra. In particular,

[a] Prof. Dr. J. Gasteiger, Dr. P. Selzer
Computer-Chemie-Centrum, Institut für Organische Chemie
Universität Erlangen-Nürnberg
Nägelsbachstraße 25, 91052 Erlangen (Germany)
Fax: (+49) 9131-85-26566
E-mail: gasteiger@ccc.chemie.uni-erlangen.de

[b] Dipl.-Chem. H. Thomas, Prof. Dr. R. Salzer
Institut für Analytische Chemie
Technische Universität Dresden
Mommsenstraße 13, 01069 Dresden (Germany)
Fax: (+49) 351-463-7188
E-mail: reiner.salzer@chemie.tu-dresden.de

0947-6539/00/0605-0920 $ 17.50+.50/0 *Chem. Eur. J.* **2000**, *6*, No. 5

we will investigate how a rather limited database of 13373 infrared spectra and their corresponding structures allows one to make predictions over a broad range of organic substances. This will point out the scope and limitations of this approach and some deficiencies of such a small database.

Simulation of IR Spectra

Neural networks: The relationship between structures and infrared spectra is modelled by an artificial neural network.[7, 8] As illustrated in Figure 1, the neural network used in this

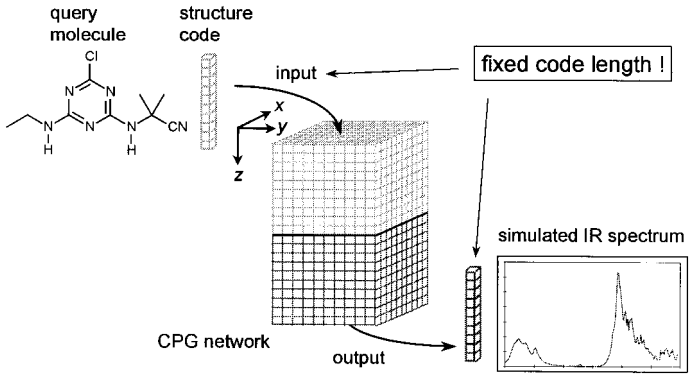


Figure 1. Simulation of an infrared spectrum by a neural network. The neural network requires a fixed length code for the input (structure representation) and the output (infrared spectrum).

approach, a counterpropagation (CPG) neural network,[9] consists of a rectangular arrangement of $x \times y$ neurons (for example 10×10). Each neuron has z weights (for example 256, 128 for the representation of the structure and 128 absorbance values for the representation of the infrared spectrum). Neural networks learn inductively; this means that they learn to model the relationship between structures and spectra by analyzing a set of examples (molecular structures and their corresponding infrared spectra) in the so-called training process. During the training the weights of the neurons are adjusted to become more similar to the training data. After training, the neural network is able to predict the infrared spectrum for a molecule the network has not seen before from the weights stored in the network. In the simulation process the Euclidean distance between the structure code of the query molecule (input) and each neuron is calculated. The neuron having the lowest Euclidean distance to the query structure is the winning neuron and supplies the predicted infrared spectrum in a lookup process as output.

Abstract in German: *Die Infrarotspektroskopie eignet sich aufgrund der hochcharakteristischen Banden sehr gut zur Substanzidentifikation. Dazu wird in der Regel das experimentelle Spektrum der zu identifizierenden Substanz mit einem Referenzspektrum aus einer Datenbank oder einem Spektrenkatalog verglichen. Ist das gesuchte Spektrum jedoch in keiner Datenbank enthalten, bleibt dieser einfache Weg der Substanzidentifikation versperrt. Eine mögliche Lösung ist, das entsprechende Referenzspektrum zu simulieren. Basierend auf neuronalen Netzen wurde eine Spektrensimulationsmethode entwickelt, die Zugang zu Referenzspektren für organische Verbindungen bietet. Die Möglichkeiten und Grenzen dieser Methode werden anhand eines Datensatzes von 16 Testverbindungen, die einen breiten Bereich an funktionalen Gruppen abdecken, diskutiert.*

Structure coding: A fundamental requirement of neural networks is that the input and the output has to be represented by a fixed number of variables. For the output this is rather straightforward: Each spectrometer represents the measured spectrum by a fixed number of variables, for example, 2000 absorbance values at certain wavenumbers. For the molecular structure a fixed length representation is not that obvious. The common representation of chemical structures by Cartesian coordinates cannot be used since, in this case, the number of variables depends on the number of atoms of the molecule. As mentioned above we have developed a method that transforms the 3D structure of a molecule into a fixed length code. In the first step, the molecular 3D structure is generated from the connection table by the automatic 3D structure generator CORINA.[10, 11] Then, physicochemical properties (for example the total atomic charges q_{tot}) are rapidly calculated by empirical methods collected in the program package PETRA.[12–14] The 3D structure is then transformed into the structure code, the so-called radialcode, while simultaneously considering these physicochemical properties (Figure 2).[15]

This radialcode $g(r)$ is calculated as given in Equation (1):

$$g(r) = F \cdot \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j e^{-B(r-r_{ij})^2} \quad (1)$$

with:
N: number of atoms
A_i, A_j: atomic property of atom i and j
r_{ij}: interatomic distance between atoms i and j
B: temperature or smoothing parameter
F: scaling factor

The code is calculated with the variable r running in discrete equidistant steps (for example 128 steps) from 0 to 12.8 Å. The resulting code is a sum of all interatomic distances in a molecule. In Figure 3 the codes for benzene, toluene, and the three xylene isomers are shown. In order to simplify the presentation, Figure 3 shows only the values calculated for the C–C distances. In the actual simulation experiments the distances between all pairs of atoms were considered in the calculation of the codes.

Since the method is based on correlating experimental data, the computational time and the prediction quality are nearly independent of the size of the molecule. The prediction quality highly depends on the data that were used for training. Two aspects are important: First, it is necessary that the training spectra are of high experimental quality. For example, if all training spectra have been taken in insufficiently dried KBr, the network would learn that each infrared spectrum has to show a broad band at 3400 cm^{-1}, because of the water contents of KBr.

And secondly, for high prediction quality it is necessary, that the query structure is well represented by the molecules of the training set. In the ideal case, the query molecule can be

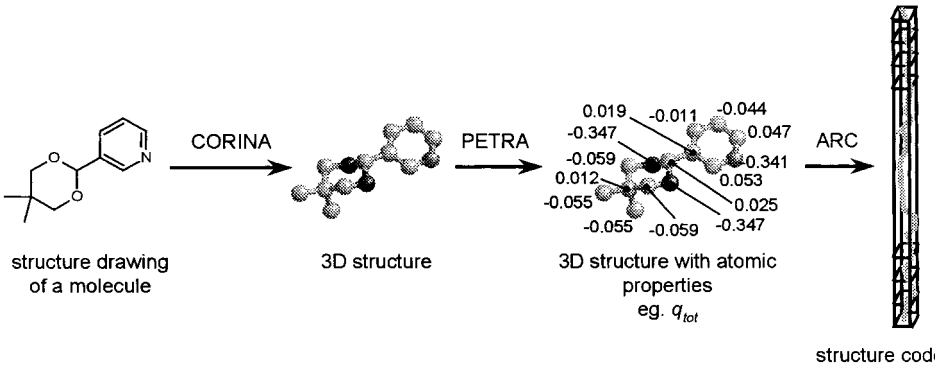


Figure 2. Transformation of a 3D molecule structure into a fixed length code. In the first step the structure drawing is converted into a 3D structure. After calculating physicochemical properties, the 3D structure is transformed into a structure code.

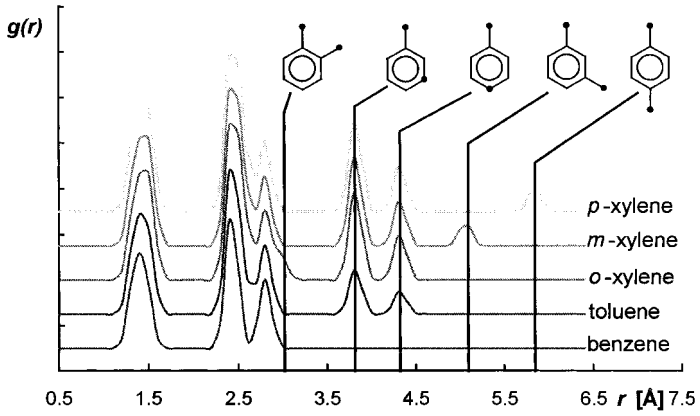


Figure 3. The radial distribution function (RDF) code for benzene derivatives. The first three peaks correspond to the 1–2, 1–3, and 1–4 C–C distances in the benzene ring.

interpolated between structures of the training set. A CPG-neural network is a very good interpolator, whereas its extrapolation abilities are poor.

For the selection of the training molecules the structure code of the query structure is compared with the structure codes of all entries in the structure and spectra database (Figure 4). The criterion for comparison is the root mean square (*rms*)-error in the values of RDF code. Those 50 molecules showing the highest similarity with the query structure code were taken into the training set.

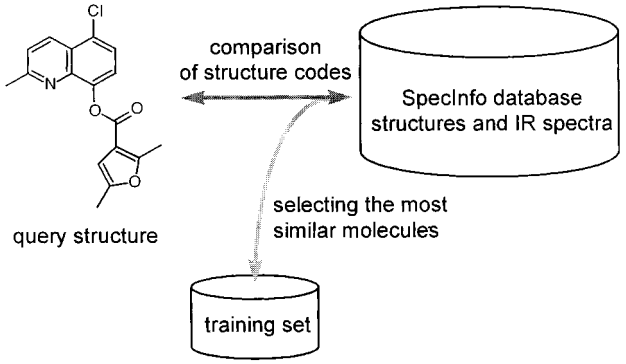


Figure 4. Query driven selection of the training set. The 50 molecules having the most similar structure code compared with the query structure were taken for the training of the neural network.

This approach offers a high potential for the adaptation to the problem, since each query molecule defines its own training set and its own specially trained network.

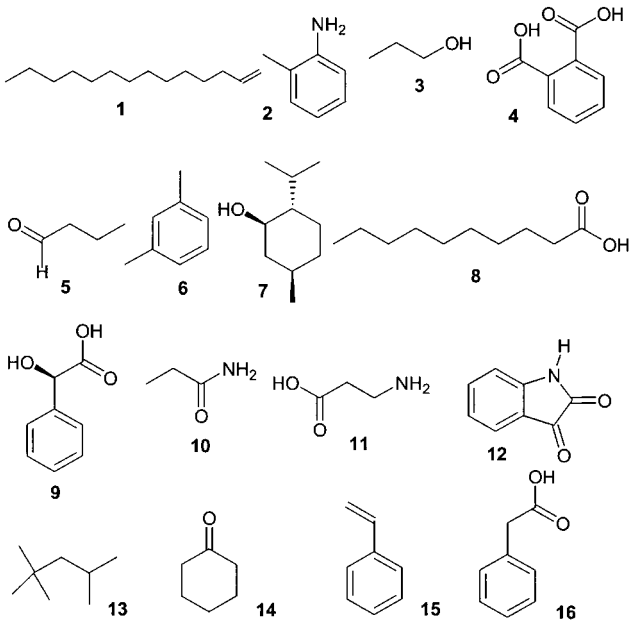
Computational Methods

In principle, the method can be applied to arbitrary compounds. The spectra that were used for training are taken from the SpecInfo[16] infrared spectra database. The aim of this experiment was to analyze if a limited database such as this with 13373 molecules supplies sufficient information to perform reasonable simulation experiments.

The test set

To test the applicability of the method, 16 compounds with a variety of structural features were selected by independent coworkers to cover a wide range of chemistry. Twelve compounds had only one functional group and can be assigned to a specific substance class, whereas four compounds contain a combination of two or more functional groups. The analysis of the simulation for these combinations of functional groups is of particular interest, since it illustrates whether the system is able to react on spectral changes that might be caused by these structural features. The 16 test compounds are shown in Scheme 1.

It must be mentioned that the spectra of compounds **1**, **2**, **3**, **6**, **7**, **8**, **9**, **13**, **14**, **15**, and **16** are included in the SpecInfo database, but these spectra were not used for the training of



Scheme 1. Molecules of the test set.

0947-6539/00/0605-0922 $ 17.50+.50/0 *Chem. Eur. J.* **2000**, *6*, No. 5

the neural network. Therefore the experiment was performed as if these compounds were unknown. Liquid substances were taken as films, waxy substances were melted. The spectra were described by 128 absorbance values between 3500 and 560 cm^{-1} with a resolution of 40 cm^{-1} between 3500 and 2020 cm^{-1}, and a resolution of 16 cm^{-1} between 2000 and 560 cm^{-1}.

Results and Discussion

A query driven infrared spectrum simulation experiment was performed for all molecules of the test set (Scheme 1). The spectrum simulation approach described here is quite fast. It takes 90 seconds on an SGI ORIGIN 200 from the input of the structure to the output of the simulated spectrum. The computation time includes the 3D structure generation, the calculation of physicochemical properties, the transformation into the structure code, the selection of the training set, the neural network training, and the spectrum prediction process. It should be emphasized that neural networks allow a separation of the more time-consuming training phase from the test phase making the latter nearly instantaneous. This can be achieved when a neural network is needed for a set of related compounds, for example, for substituted benzene, or quinoline compounds. Here, however, we decided to use the query directed approach that merges training and test phase in order to explore which structures are included in the database that are similar to each individual molecule of the test set. The simulation quality was determined by calculating the correlation coefficient r,[17, 18] between simulated and experimental spectra. For a better visual comparison of the simulated and the experimental spectra the absorbance values of the experimental spectra were adjusted by setting the lowest value equal to 0, which has no effect on the correlation coefficient r. [It has to be emphasized that the symbol r is used for the correlation coefficient as is standard use. However, this use of r has to be distinguished from the distance variable r used in Eq. (1).]

From the spectroscopic point of view the correlation coefficient r has weaknesses since it is a statistical measure that does not emphasize the position and the relative intensities of important bands. Analyzing the simulation experiments and the training data we observed that experimental spectra from the training set have a higher correlation coefficient with the query spectrum than the simulated spectrum, even if the correspondence of bands is similar. So the simulated and the experimental spectrum still have to be compared by visual inspection to analyze how good the important bands and the overall shapes were reproduced. This will be discussed in detail with some simulation examples. However, the correlation coefficient r is quite a reasonable similarity measure for the comparison of IR spectra since it is less sensitive to differences in absolute intensities than, for example, the *rms* error. Figure 5 shows the distribution of the calculated correlation coefficients for the 16 simulation experiments.

It can be observed that in the majority of cases the simulated spectra show high similarity with the experimental data: six spectra with $r > 0.9$, six spectra with $0.9 > r \geq 0.8$, two spectra with $0.8 > r \geq 0.7$; however, there are also two spectra with $r < 0.2$.

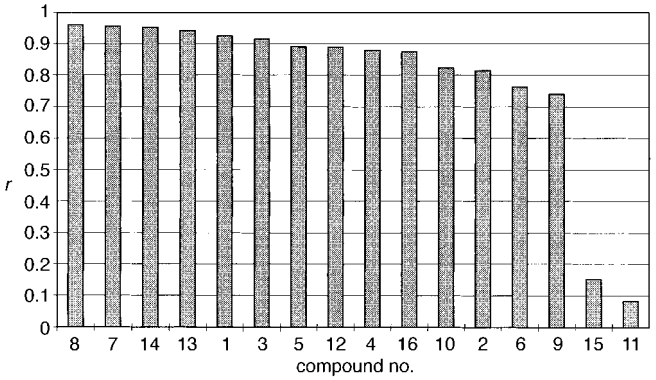


Figure 5. Distribution of the correlation coefficient r of the simulated versus the experimental spectra.

The experiments showing very high and very low correspondence with the experimental data will be discussed in more detail. Figure 6 shows the simulation for decanoic acid

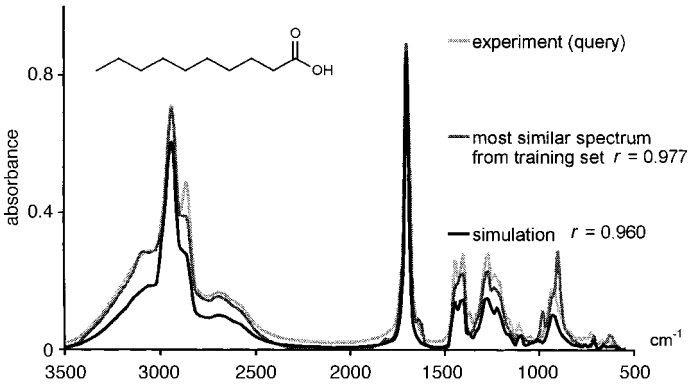


Figure 6. Comparison of simulated and experimental spectrum of compound **8**. In addition, the training spectrum having the highest correlation coefficient with the experimental spectrum is displayed, i.e., 10-undecenoic acid.

(compound **8**) having the highest correlation coefficient $r = 0.960$. In addition the training spectrum having the highest correlation coefficient ($r = 0.977$) with the experimental spectrum of compound **8** is displayed. This is the spectrum of 10-undecenoic acid.

Simulated and experimental spectrum of compound **8** show very high similarity. Even the signals in the fingerprint region are reproduced very well. The reason for this high quality simulation result is that the query structure is quite well represented by the molecules in the training set. A detailed analysis of the neural network will illustrate this. As shown in Figure 1 the neural network has a square arrangement of 10×10 neurons. In the approach described in this publication the neurons of the network are connected in a toroidal manner,[7] so the neurons in row 10 are connected to the corresponding neurons in row 1. In the same way the neurons in column 10 are connected to the corresponding neurons in column 1. Therefore each neuron has eight neurons in its first neighborhood sphere. Figure 7 shows a view on the top of the neural network and the structures that were assigned to certain

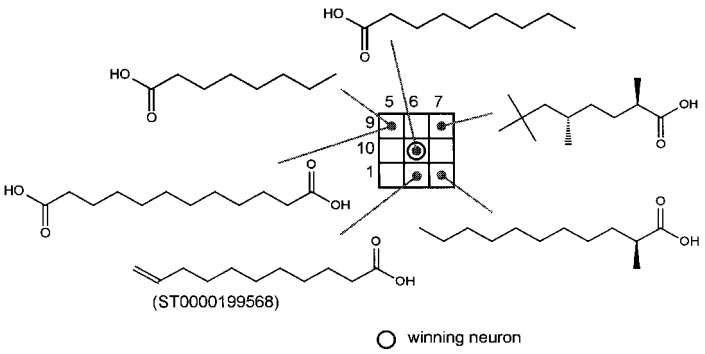


Figure 7. Detail of the neural network showing the winning neuron and the first sphere of neighboring neurons with the compounds from the training set.

neurons during training. It displays a zoomed part of the neural network with 3×3 neurons containing the winning neuron in the center (marked with a gray circle) and its first sphere of neighboring neurons.

The winning neuron contains nonanoic acid, which is very similar to the query structure, that is decanoic acid, with only one additional CH_2 group. Since the query structure contains eight CH_2 groups, this difference has only a minor effect on the shape of the spectrum. Although the molecules in the first neighboring neurons show more structural deviations to the query structure, they are nevertheless aliphatic carboxylic acids and can therefore contribute to a high quality simulation. Only the molecule in neuron 1,5 (10-undecenoic acid, SpecInfo ID ST0000199568) with one C=C double bond has an additional functional group. It is interesting to mention that in spite of this structural deviation this spectrum has the highest correlation coefficient with the query spectrum of all training spectra and the simulation result. This stresses again the importance of the visual inspection of simulation results. The simulation for (−)-menthol (compound **7**, see Figure 8) is also of high quality, as indicated by a correlation coefficient of $r = 0.955$.

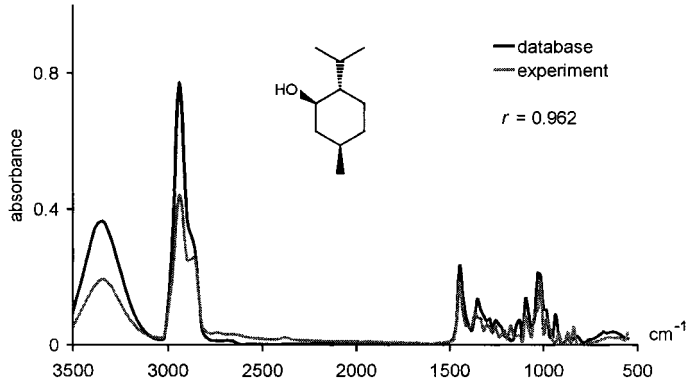


Figure 8. Comparison of simulated and experimental spectrum of (−)-menthol (compound **7**).

Again, the simulated and the experimental spectrum show high correspondence. Even the shape of the fingerprint region is very well reproduced. As in the example of compound **8** the reason for the high quality simulation is that the query structure is well represented by the molecules of the training set. All molecules in the neurons directly adjacent to the winning neuron have a cyclohexane ring (in one case a decaline system) and an OH group.

The correlation coefficient $r = 0.739$ for the simulation of compound **9** (Figure 9) is more at the lower end of the similarity scale. The simulated spectrum and the spectrum

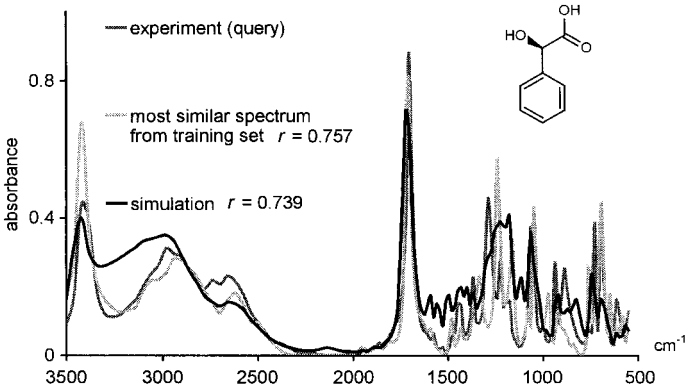


Figure 9. Comparison of simulated and experimental spectrum of DL-mandelic acid (compound **9**). In addition, the training spectrum having the highest correlation coefficient with the experimental spectrum of DL-mandelic acid is displayed, i.e., benzilic acid.

from the training set having the highest correlation coefficient with the query spectrum (benzilic acid, SpecInfo ID ST0000199520) show a similar correspondence of their signals with the signals of the query spectrum. Simulated and experimental spectrum still correspond in their shape. The largest deviations can be observed in the fingerprint region. An analysis of the molecules that have been used for the training of the neural network gives an explanation for this result (Figure 10).

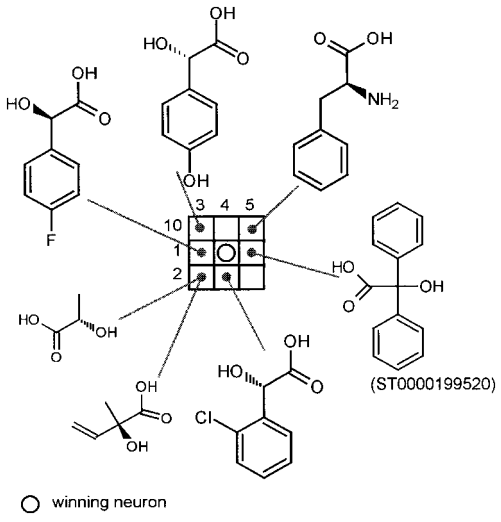


Figure 10. Detail of the neural network showing the winning neuron and the first sphere of neighboring neurons with the compounds from the training set.

The training molecules that have been assigned to the neurons in the first neighboring sphere of the winning neuron show structural similarity with the query structure. All

0947-6539/00/0605-0924 $ 17.50+.50/0 Chem. Eur. J. 2000, 6, No. 5

molecules contain a carboxyl group and a phenyl substituent except for the molecules in neuron 2,3. However, some molecules also have structural features that cannot be found in the query structure, for example, acyclic structures, halogen, or amino substituents, leading to deviations in the fingerprint region which cause a correlation coefficient r of only 0.739. Furthermore, it should be noted that the infrared spectrum was that of the racemic compound **9**, whereas for the structure input and 3D structure generation a specific stereoisomer had to be chosen.

Finally, the two examples showing very low simulation qualities will be discussed in more detail. The simulation experiment for compound **11** β-alanine basically represents no similarity to the experimental spectrum (Figure 11).

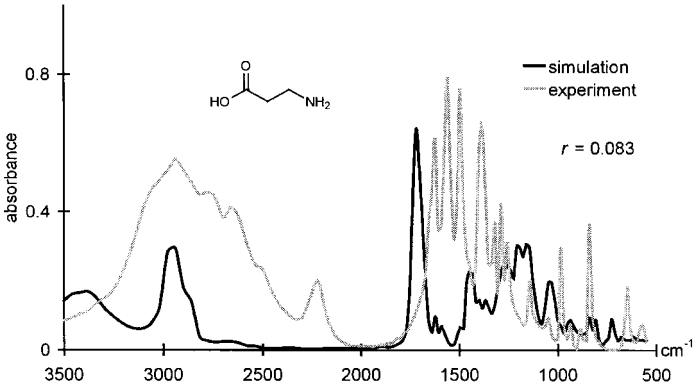
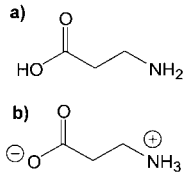


Figure 11. Comparison of simulated and experimental spectrum of β-alanine (compound **11**).

The reason for this low simulation quality will become more obvious from an analysis of the molecules that have been used for the training of the corresponding neural network. The neurons in the first neighboring sphere to the winning neuron do not contain any amino acids (Figure 12). The molecules have hardly any important structural features in common with the query structure. This indicates that the error does not occur in the step of the neural network prediction but in the step of the training set selection. The system was not able to select appropriately similar molecules for training. It has to be realized that the structure of β-alanine as in Scheme 2 used for the simulation experiment reported in Figure 11 and 12 is basically not correct, because β-alanine exists as zwitterion.



Scheme 2. a) β-alanine as neutral molecule and b) as zwitterion.

Thus, it is interesting to investigate whether the simulation quality improves if the query structure is submitted as zwitterion. Figure 13 shows the result of this experiment. The simulated and the experimental still show fairly large deviations but the correlation coefficient of $r = 0.444$ indicates a much higher correspondence between the two spectra than in the simulation before (cf. Figure 11). An analysis of the training compounds used for training, particularly those in the vicinity of the winning neuron (Figure 14) gives an explanation for this increased simulation quality: In the experiment in which the amino acid was submitted as zwitterion, the system found other amino acids in the database and used them for training the network. Since these database entries were stored as zwitterions, they had not been recognized as being similar in the experiment before, where the query amino structure was submitted with neutral amino and carboxylic group Figure 11 and 12).

This experiment underlines the importance of an appropriate description of a query structure. The system can only

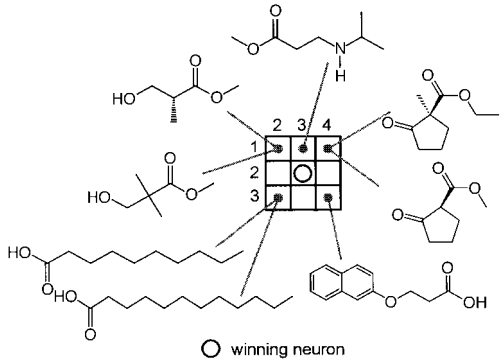


Figure 12. Detail of the neural network showing the winning neuron and the first sphere of neighboring neurons with the compounds from the training set of query compound β-alanine (compound **11**)

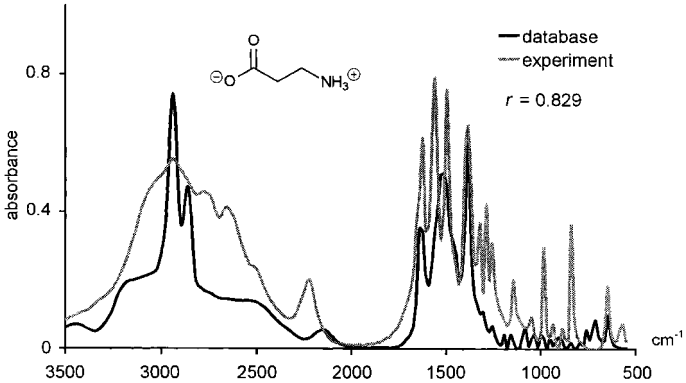


Figure 13. Comparison of simulated and experimental spectrum of β-alanine (compound 11) as zwitterion

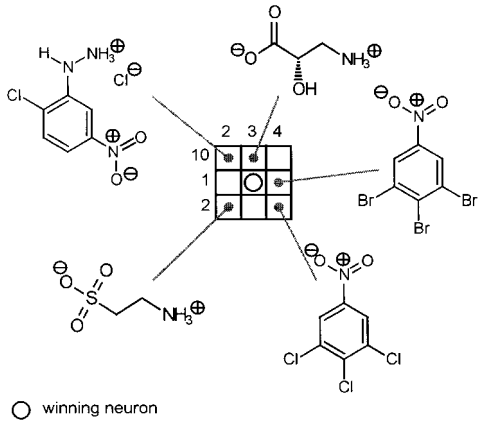


Figure 14. Detail of the neural network showing the winning neuron and the first sphere of neighbors with the compounds from the training set of query compound β-alanine (compound 11) as zwitterion

recognize structural similarity if the submitted compound is described in the same way as the compounds in the database.

The next experiment discussed in detail is the spectrum simulation for styrene (compound **15**). Again, the simulated and the experimental spectrum show very low correspondence Figure 15) with a correlation coefficient $r = 0.153$.

An analysis of the neural network supplies further information. Figure 16 shows the winning neuron (marked with a circle) and two neighboring spheres. All of the displayed

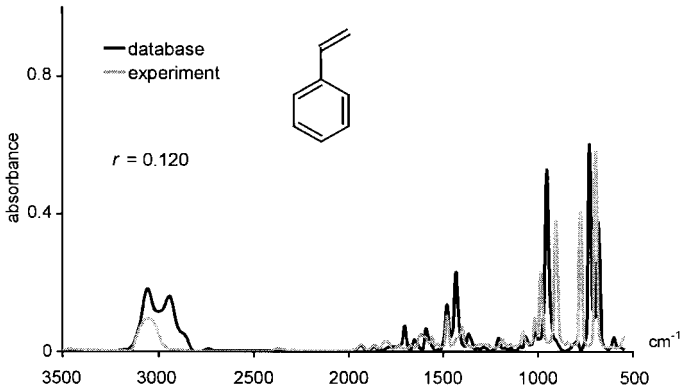


Figure 15. Comparison of simulated and experimental spectrum of styrene (compound 15)

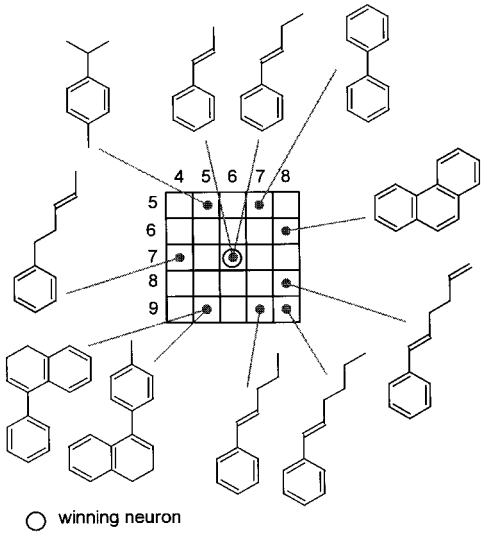


Figure 16. Detail of the neural network showing the winning neuron and the first two spheres of neighboring neurons with the compounds from the training set of query compound styrene (compound **15**).

training molecules show structural features that can also be found in the query structure: a benzene ring and an alkene group. The two training molecules from the winning neuron show high similarity with the query structure: Whereas styrene has a benzene ring substituted with a vinyl group, one training structure is substituted with a 1-propenyl and another one with a 1-butenyl group. These rather small structural deviations, however, cause an enormous effect from the spectroscopic point of view, because infrared spectroscopy reacts very sensitive whether or not a compound has only olefinic or aliphatic structural features. The additional CH_2 groups in the two training molecules cause quite a few bands that are absent in the query structure. The rule to be learnt is that in this approach, structures with only a few features that are active in an infrared spectrum will usually lead to a lower simulation quality. By analyzing the training structures any spectroscopist could estimate that this particular simulation result might not be reasonable.

Internet access: This spectrum prediction method can be accessed through the internet (http://www2.ccc.uni-erlangen.de/IR/). This web page features interactive simulation experiments and database searches. The simulation experiments described in this article are stored on the server and can be retrieved with the keyword "public" following the open-link at the page http://www2.ccc.uni-erlangen.de/IR/simuframe/.

Summary and Conclusions

Infrared spectroscopy is, as a result of its highly characteristic bands, very useful for the substance identification by comparing the experimental spectrum with the reference spectrum from a database. Because of the very unfavorable relation between the amount of 16000000 known compounds and the number of only 100000 infrared spectra in the largest IR spectra database, this easy way of substance identification often fails. Therefore, there is a need for spectrum prediction methods to close these data gaps. In this article, a spectrum prediction method based on neural network techniques was presented that provides rapid access to arbitrary reference spectra. Since the method is based on experimental data we investigated if a database with 13373 infrared spectra supplies sufficient spectral information to perform reasonable prediction experiments.

The experiments reported herein have shown that the presented spectrum prediction method can provide reasonable prediction results for a broad range of organic compounds. From a test set of 16 compounds selected by independent scientists, six spectra show a correlation coefficient $r > 0.9$ between simulated and experimental spectrum and therefore give very high correspondence. An additional six simulated spectra have a correlation coefficient of $0.9 \geq r > 0.8$ indicating high similarity. Two spectra show a correlation coefficient of $0.8 \geq r > 0.7$ which still displays acceptable similarity between simulation and experiment. Only two spectra give an $r < 0.2$ which indicates poor similarity. The reason for this low correspondence could be clarified. Since the prediction method is based on data, the simulation quality is the higher the higher the similarity between the query structure and the training structures is. Therefore, important information can be derived from analyzing the compounds that have been used for the training of the neural network. This information allows the user to estimate how reliable the prediction experiment is. The overall success of this method for IR spectra prediction attests to the power of the radial distribution function to code important 3D information on molecules.

0947-6539/00/0605-0926 $ 17.50+.50/0 *Chem. Eur. J.* **2000**, *6*, No. 5

Acknowledgments

Support of this work by the Bundesminister für Bildung, Wissenschaft, Forschung und Technik is gratefully acknowledged. We thank the Verein zur Förderung des Deutschen Forschungsnetzes e.V. DFN, for funding the TeleSpec-Project and Chemical Concepts for providing us with the SpecInfo database. The authors wish to thank Markus Hemmer for many stimulating discussions about the structure coding and his help in preparing this paper.

[1] P. Selzer in *Encyclopedia of Computational Chemistry* (Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. Schreiner), Wiley, Chichester, UK, **1998**, p. 1299–1306, and references therein.
[2] C. Affolter, J. T. Clerc, *Lab. Inf. Magn.* **1993**, *21*, 151–157.
[3] U.-M. Weigel, R. Herges, *Anal. Chim. Acta* **1996**, *331*, 63–74.
[4] J. H. Schuur, P. Selzer, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.
[5] J. H. Schuur, J. Gasteiger, *Anal. Chem.* **1997**, *69*, 2398–2405.
[6] M. C. Hemmer, J. Gasteiger, *Vib. Spectrosc.* **1999**, *19*, 151–164.
[7] J. Zupan, J. Gasteiger, *Neural Networks for Chemists—An Introduction*, VCH, Weinheim, **1993**, J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, 2nd ed., Wiley-VCH, Weinheim, **1999**.
[8] J. Gasteiger, J. Zupan, *Angew. Chem.* **1993**, *105*, 510–536; *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503–527
[9] R. Hecht-Nielsen, *Applied Optics* **1987**, *26*, 4979–4984.
[10] J. Sadowski, J. Gasteiger, G. Klebe, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.
[11] Access to the CORINA program through the Internet: http://www2.ccc.uni-erlangen.de/software/corina/
[12] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.
[13] M. G. Hutchings, J. Gasteiger, *Tetrahedron Lett.* **1983**, *24*, 2537–2540.
[14] Access to the PETRA program package through the Internet: http://www2.ccc.uni-erlangen.de/software/petra/
[15] J. Gasteiger, J. H. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, *Fresenius J. Anal. Chem.* **1997**, *359*, 50–55.
[16] SpecInfo Infrared Database, Chemical Concepts, Weinheim, Germany.
[17] G. E. P. Box, W. G. Hunter, J. S. Hunter, *Statistics for Experimenters*, Wiley, New York, USA, **1978**.
[18] J. Bortz, *Statistik für Sozialwissenschaftler*, 3rd ed., Springer Verlag, Berlin **1989**, p. 213.

Received: May 21, 1999 [F1800]