Chimia 52 (1998) 678–682
© Neue Schweizerische Chemische Gesellschaft
ISSN 0009–4293

IR Spectra Simulation and Information Processing on the WWW

Paul Selzer*

Abstract. Substance identification by IR-spectroscopic methods generally is performed by comparing the measured spectrum with a reference spectrum from a database. If the desired spectrum is not contained in the database, another method of substance identification has to be found.

Based on neural network techniques, we have developed a method which provides rapid access to simulated reference spectra. Within the scope of the 'TeleSpec-Telecooperation in Spectroscopy'-Project, this algorithm has been made available through a WWW interface. With a JAVA-enabled Web browser, interactive spectra prediction experiments can be performed.

Introduction

Due to the highly characteristic bands and the broad range of sample preparation techniques, IR spectroscopy is a very useful method in analytical chemistry.

Substance identification by IR-spectroscopic methods is usually performed by comparing an experimental spectrum with a reference spectrum from a spectrum library. This identification technique requires that a reference spectrum for the query spectrum is available. The high discrepancy between the amount of more than 16000000 known chemical compounds and only 100000 spectra stored in the largest IR spectra database often prevents this easy way of substance identification. In this paper, we present a method which is based on a combination of a neural network with a novel structure, coding scheme that allows the rapid simulation of IR spectra and thus supplies access to reference spectra for nearly arbitrary query molecules [1].

Within the scope of the 'TeleSpec-Telecooperation in Spectroscopy'-project, this method for spectrum prediction and interpretation has been implemented as a freely accessible Web service (http://www2.ccc.uni-erlangen.de/IR/). We consider this project to be an advanced example and case study for the current general trend towards presentation and interfacing of chemical information processing on the WWW.

Simulation of IR Spectra with Neural Networks

The correlation between the structure of a molecule and its IR spectrum is very complex and cannot be described by a simple formula. Neural networks are very useful tools for the modeling of such nonlinear relationships [2]. An important advantage of neural networks is that they learn inductively. This means that they are capable of learning about the relationship between structure and spectrum by analyzing a set of examples. The neural network acts as an interpolator. So no *apriori* knowledge has to be fed into the system. This method is based on experimental data, it is very fast, and can be applied to arbitrary molecules, provided that a suitable body of previously measured structures of comparable characteristics is available. The computation time and the prediction quality is nearly independent of the size of the molecule. For a high prediction quality, it is required that the training set contains similar molecules and that the experimental spectra that were used for training are of high quality. These aspects will be explained in more detail in the following chapters.

Presenting Molecule Structures to the Neural Network

A fundamental requirement of neural networks is that the input and the output have to be described by a fixed number of variables. For the output, the IR spectrum, this is obvious. In fact, every spectrometer supplies a representation of a spectrum by a fixed number of variables, *e.g.*, 2000 absorbance values at certain wave numbers. The representation of the molecular structure in this fashion is a more complex task. Customary structure descriptions, *e.g.*, cartesian atom coordinates, cannot be used, because they depend on the number of atoms of the molecule. Based on the fact that IR spectroscopy monitors the vibrations of atoms and groups in three-dimensional space, we have developed a method for the encoding of the three-dimensional structure of an entire molecule which satisfies the requirements outlined above [3]. First, the structure drawing of the query molecule, stored as connection table, is automatically converted into a 3D structure with the aid of a model builder. Then, physicochemical properties are calculated, *e.g.*, the atomic mass m, the local polarizability α, or the total atomic charge q_{tot} [4][5]. Taking these physicochemical properties into account, the 3D structure is transformed into a structure code (*Fig. 1*).

The resulting structure code, called Radial Distribution Function (RDF) Code, is calculated from internal coordinates exclusively and is therefore rotation- and translation-invariant.

The *Eqn.* for the calculation of the Structure code

$$g(R) = f\sum_{i}^{N-1}\sum_{j>i}^{N} A_i A_j \cdot e^{-B(R-R_{ij})^2}$$

with:

R: interatomic distance; the code is usually calculated for $R = 0 \ldots 12.8$ Å
A: atomic property, *e.g.*, the atomic mass m, the polarizability α, or the total atomic charge q_{tot}
B: temperature or smoothing parameter
N: number of atoms
f: scaling factor

The inclusion of the physicochemical properties in the computation increases the information content of the structure code with respect to the IR spectrum. This is of eminent importance since the method is based on the analysis of the correlation between this code (a complex way of encoding a chemical structure) and the IR spectrum, which is considered to be just another special description of a molecular structure.

**Correspondence*: Dr. P. Selzer
Computer Chemistry Center
Institute of Organic Chemistry
University of Erlangen-Nürnberg
Nägelsbachstrasse 25
D–91052 Erlangen
E-Mail: paul.selzer@ccc.chemie.uni-erlangen.de

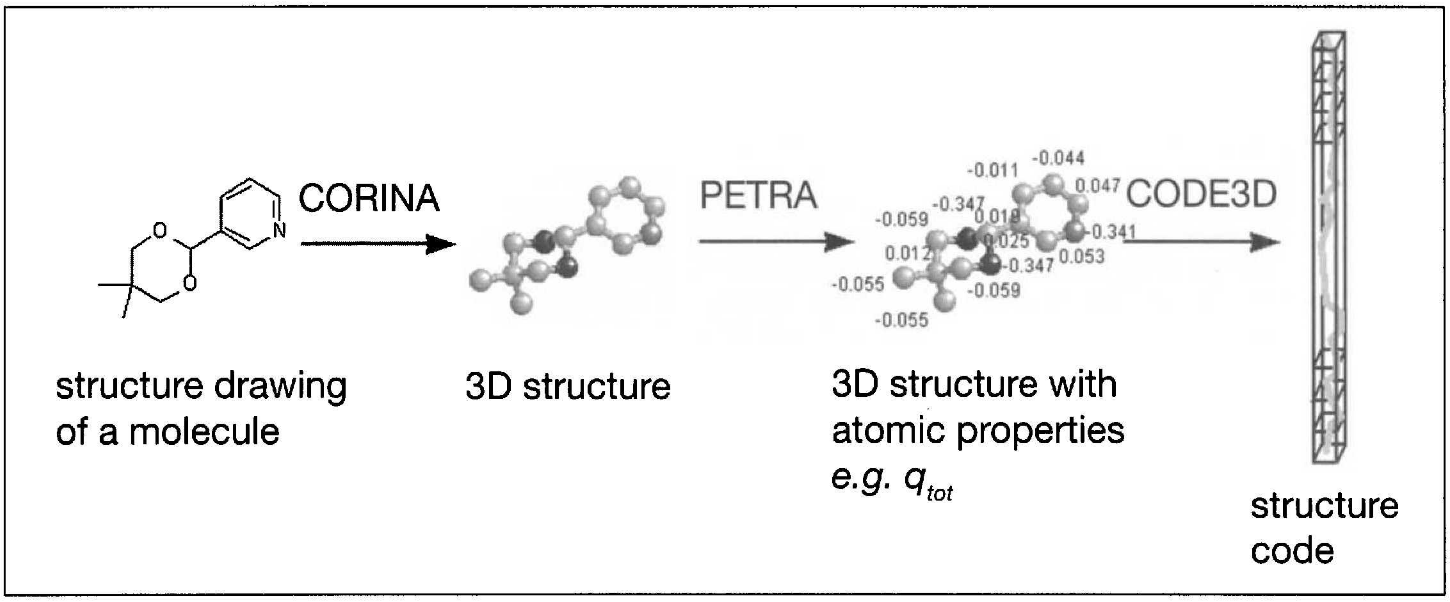


Fig. 1. *Transforming the structure into the structure code*

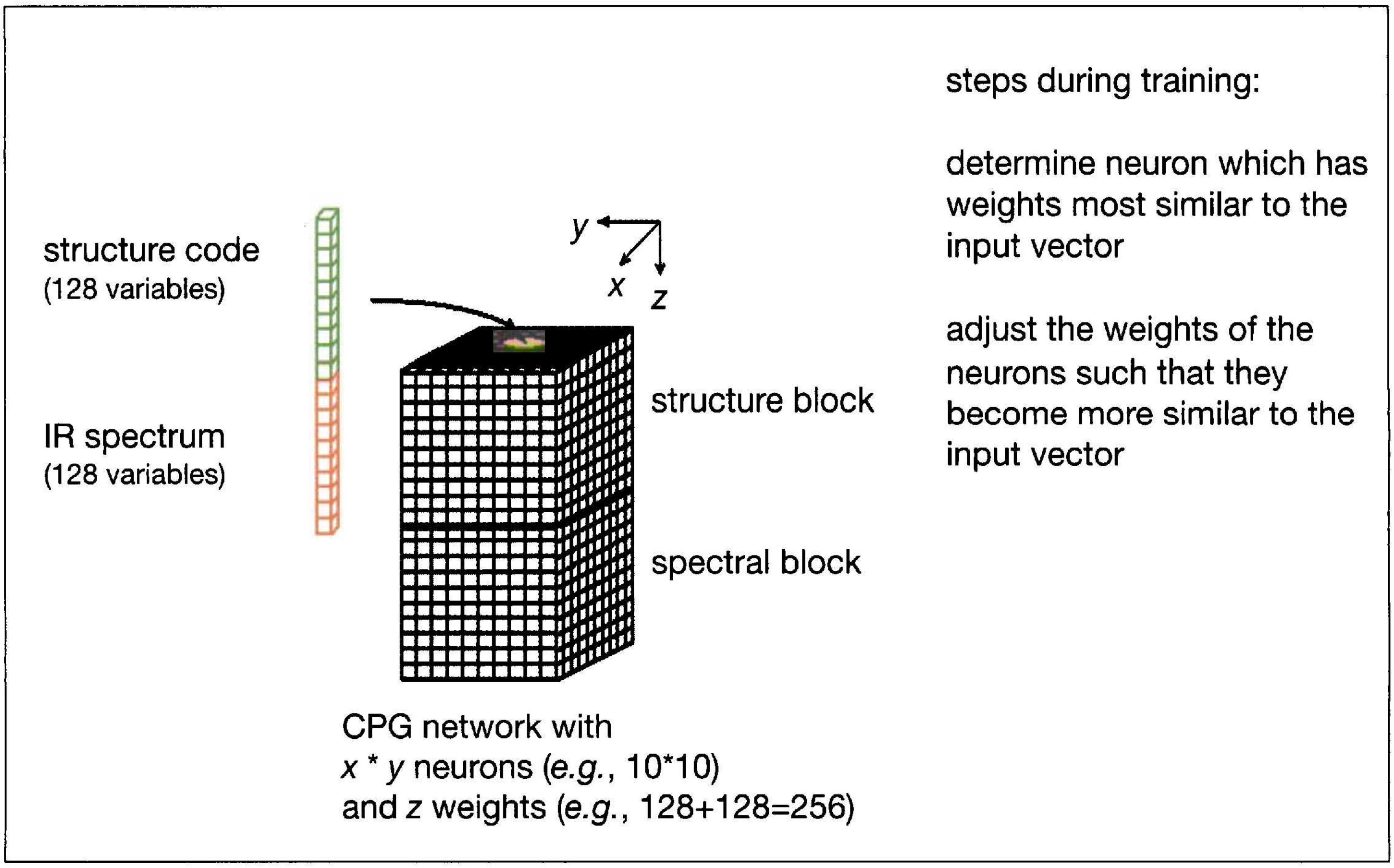

steps during training:
determine neuron which has weights most similar to the input vector
adjust the weights of the neurons such that they become more similar to the input vector
structure code
(128 variables)
IR spectrum
(128 variables)
y
x
z
structure block
spectral block
CPG network with
x * y neurons (e.g., 10*10)
and z weights (e.g., 128+128=256)

Fig. 2. *The training of a neural network consisting of two repeated basic steps: first, determination of the most similar neuron, and then, adjustment of the weights of the network*

The Network Training

As mentioned before, the neural network learns about the relationship between structure and spectrum by analyzing a set of examples. The set of examples consists of pairs of structure codes and their corresponding IR spectra [6]. These structure code/spectrum pairs are called data points. In the training process, the weights of the neural network neurons, which are initialized to random values, are adjusted in such a way that they become more similar to the training data (*Fig. 2*).

The data is represented on the network neurons as fixed-length vectors consisting of a structure-encoding subvector and a spectral information part. In the first step of the network training, for a data point, the most similar neuron is determined by calculating the *Euclid*ean distance between the structure code of a training data point and the structure block of the data vector at each neuron. The neuron having the lowest *Euclid*ean distance in this vector comparison is regarded the winning neuron. The weights (vector elements) of this neuron are adjusted to become more similar to the training data point. The weights of the other neurons are also adjusted; the degree of adjustment of a neuron decreases with increasing distance of this neuron from the winning neuron. Since the method is based on the interpolation of experimental data, the prediction is highly dependent on which data have been used for training. In out current implementation, we routinely take 50 molecules from the SpecInfo IR database, selected for exhibiting the most similar structure code to the query structure by a linear database scan. During training, these 50 molecules are presented to the network several times. After the training, the network has learned about the relationship between structures and spectra and is now able to predict a spectrum for a similar structure it has not seen before.

The Spectrum Simulation

The first step of the prediction is very similar to the training: for the structure code of the query structure, the most similar neuron is determined. Again, this is performed by the calculation of the *Euclid*ean distance between the structure code of the query structure vector and the structure part of the corresponding neuron vectors. Opposite to the training phase, no adjustment of the weights is performed now. The spectrum part of the data vector of the winning neuron supplies the predicted spectrum. *Fig. 3* shows the result of a spectrum simulation experiment for a quinoline derivative.

The simulated and the experimental spectrum exhibit a high similarity. Such results can be achieved if the general structural class and characteristics of the query structure is well represented by the molecules in the training set.

Scope and Limitations

Since the method is based on the correlation of entire 3D structures and full spectra, it allows the prediction of the complete shape of IR spectra. It is not limited to the prediction of certain signals in correspondence to the presence or absence of substructure fragments. The network learns inductively about the relationship between structures and spectra, so that no *a priori* knowledge has to be fed into the system. The method is very fast in comparison to other computational approaches. It takes only about 90 seconds from the input of a query structure to the output of the simulated IR spectrum on a moderately powerful compute server, a time span which makes this method suitable even for interactive applications. The computation time and the prediction quality is nearly independent of the size of the molecule. For a high prediction quality, it is necessary that the training set contains structures that are similar to the query structure and that the query structure can be interpolated. In general, extrapolation abilities of neural networks are poor. For example, the simulation for methane will not lead to a reasonable result, since it has to be extrapolated from the most similar aliphatic compound, ethane. This example, this must be clearly said, is not the typical application for this method. Unfortunately for a public presentation of a new computational algorithm, such structures which are too simple for our method are often input first as test cases by vaguely interested first-time users. The method works very well for larger molecules with lots of functionality, if the database contains suitable precedence structures.

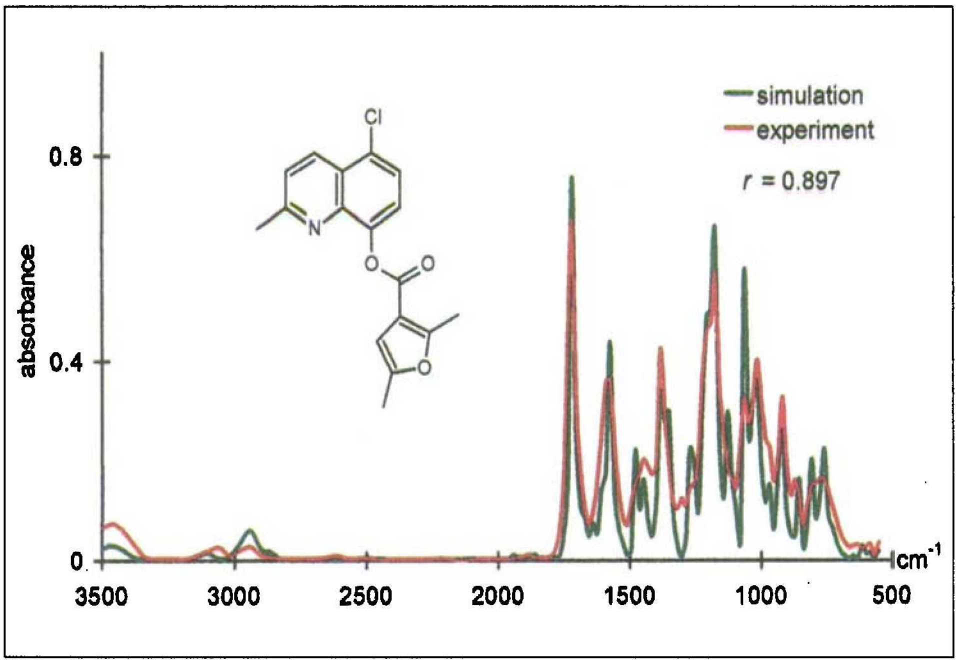


Fig. 3. *A result of a spectrum simulation experiment for a quinoline derivative.* The simulated and the experimental spectrum exhibit high similarity.

And last, this is obvious, the prediction can only be as good as the data that have been used for the training of the network. In other words: if all experimental data have been taken in moist KBr, the network learns that every iIR spectrum has to show a broad band at 3400 cm^{-1}. The higher the experimental quality of the training spectra and the larger the database, the better the results of the prediction are.

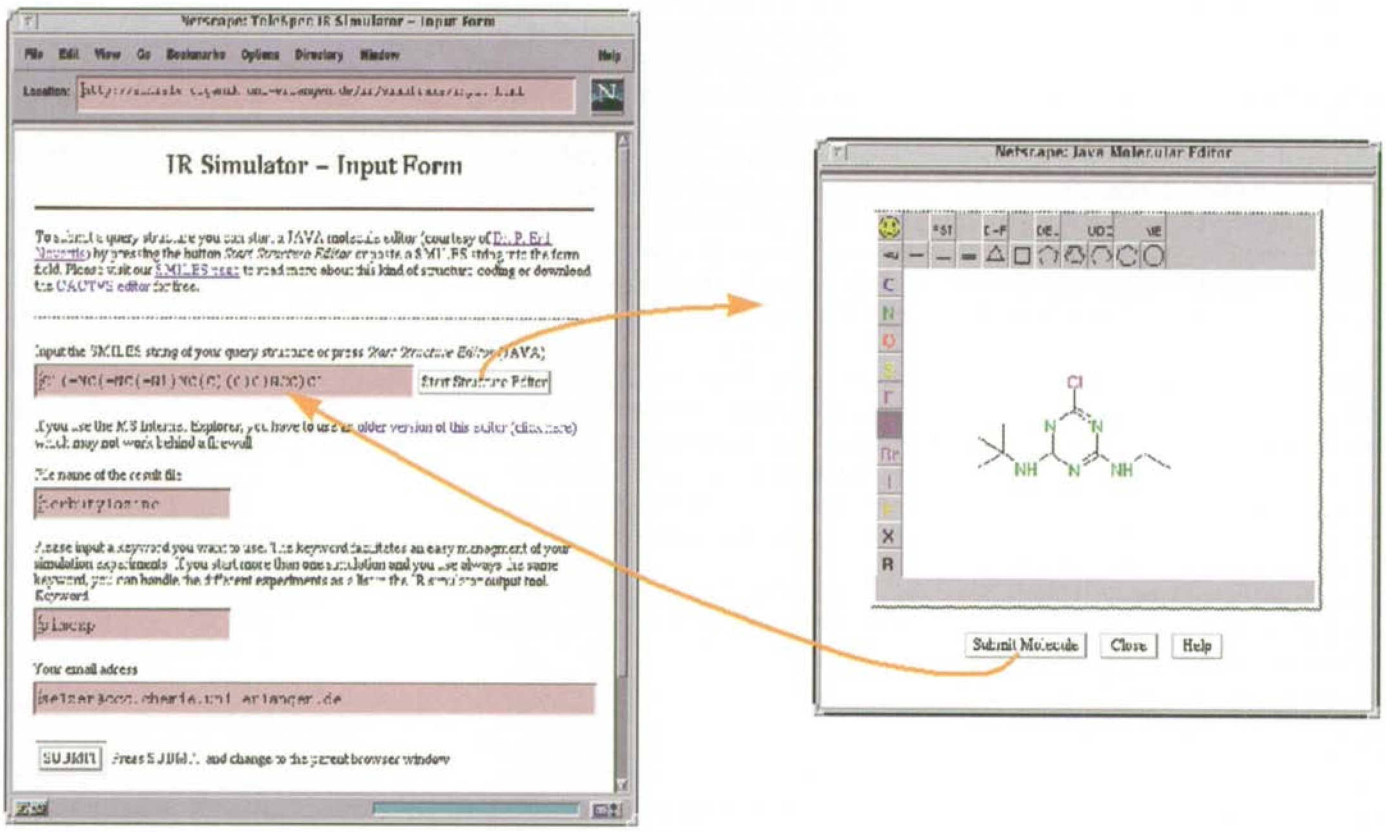
Interactivity through the Internet – The TeleSpec Project

The aim of the 'TeleSpec-Telecooperation in Spectroscopy'-Project is to provide access to this spectrum prediction method through a WWW interace and to establish an Internet-based spectra discussion forum. With a JAVA-enabled Web browser, any user can perform interactive spectrum prediction experiments, analyze and download the results. An analyst who cannot find the desired reference spectrum in his database or spectrum catalogue can connect to the TeleSpec server (http://www2.ccc.uni-erlangen.de/IR/) and simulate it.

Submitting a Query Molecule

In the first step, the user has to submit the query molecule for which he wants to run a spectrum simulation experiment. The input of a query structure is performed in the widely used Daylight SMILES string format. SMILES is a comparatively simple ASCII encoding of the connectivity and the stereochemistry of the molecular structure [7]. The SMILES code can be generated with a locally installed molecule editor and pasted into the corresponding input field. If no familiar editor with this capability is available, the well-known JAVA molecule editor by *Ertl* [8] (and in this *Chimia* issue) can be launched from the Web page (*Fig. 4*).

The user is asked to fill in a name for the result file and to select an arbitrary keyword. If the user visits the TeleSpec pages again, the keyword allows him/her to retrieve results from experiments started at former sessions, for renewed analysis or previously generated data, or if he/she did not want to wait for the results when the computations were submitted. The results are currently stored for two weeks. If several spectrum prediction experiments were performed, it is advisable to use the same keyword to group the data sets.



Netscape: TeleSpec IR Simulator – Input Form
File Edit View Go Bookmarks Options Directory Window Help
Location:

IR Simulator – Input Form

To submit a query structure you can start a JAVA molecule editor (courtesy of Dr. P. Ertl Novartis) by pressing the button *Start Structure Editor* or paste a SMILES string into the form field. Please visit our SMILES page to read more about this kind of structure coding or download the CACTVS editor for free.

Input the SMILES string of your query structure or press *Start Structure Editor* (JAVA):

C1(=NC(=NC(=N1)NC(C)(C)C)NCC)Cl

Start Structure Editor

If you use the MS Internet Explorer, you have to use an older version of this editor (click here) which may not work behind a firewall

The name of the result file

terbutylazine

Please input a keyword you want to use. This keyword facilitates an easy management of your simulation experiments. If you start more than one simulation and you use always the same keyword, you can handle the different experiments as a list in the IR simulation output tool.
Keyword

simexp

Your email adress

selzer@ccc.chemie.uni-erlangen.de

SUBMIT Press SUBMIT and change to the parent browser window

Fig. 4. *The input of a query structure in SMILES string format.* The SMILES string of the input molecule can be generated with a locally installed molecule editor or a JAVA molecule editor that can be launched directly from the Web page.

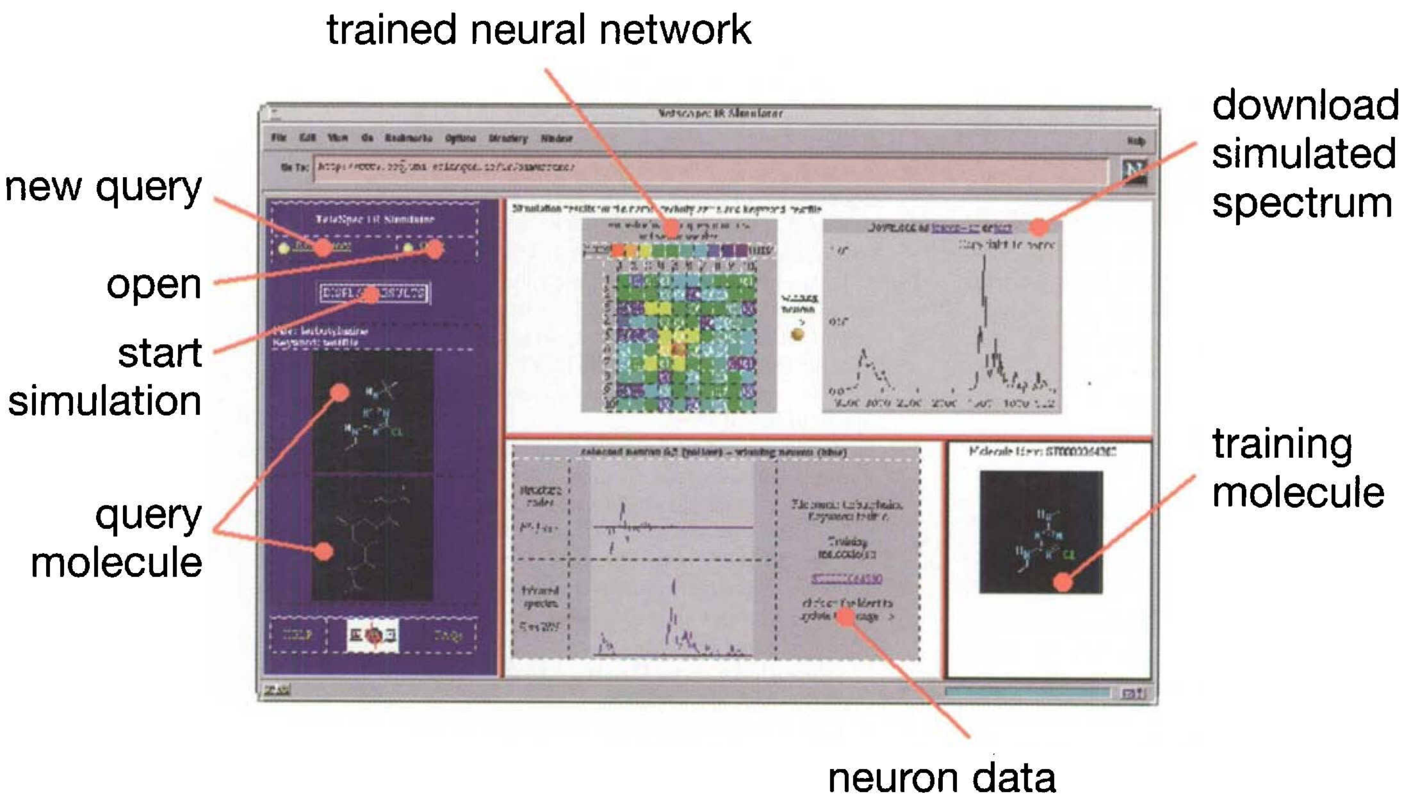


Fig. 5. *The prediction result downloaded in JCAMP-DX or tabular format.* By clicking at the individual neurons of the neuron map, the corresponding training molecules are displayed.

After filling in the input fields, the user only finally needs to press the 'SUBMIT' button and return to the simulator page.

Starting a Simulation Experiment

At the entry of the simulator page set, the 2D and 3D structure of the query molecule are displayed. The computed 3D structure is displayed in a JAVA applet of the ChemSymphony suite (see paper of *A. Krassavine* [9] in this issue) with a rotatable and zoomable stick model of the query molecule [10].

After pressing 'START SIMULATION', the spectrum prediction experiment is launched. After *ca.* 90 seconds, the computation finishes and the results become available. The right side of the window shows the simulated spectrum. The user can download it as JCAMP-DX [11] file or in a tabular text format. However, a number of tools for analysis are available directly from the Web page.

For example, we provide a clickable depiction of the trained neural network. The colors of the neuronal map indicate the similarity between the structure code of the query molecule and the weights of the neurons. Neurons which are marked with a benzene ring have been assigned one or more molecules during the training process. The example shown in *Fig. 5* can be retrieved through the TeleSpec pages using the keyword 'chimia'.

Analysis of the Neural Network

The assignment of molecules in the training is a very important information, since it allows the user to estimate the prediction quality. By clicking at the neurons, the corresponding training structures are displayed in the lower frame of the window. The user can check whether these training molecules selected from the limited pool of the underlying database structures are similar enough to the query molecule or not. As mentioned before, the prediction quality is higher as better the query molecule is represented by the training molecules.

Conclusions

The described IR spectra prediction method can be applied to arbitrary molecules. It opens new venues in substance identification and IR spectra interpretation. Since the method is based on the correlation of structural information with experimental spectra, the short computation time and the prediction quality are nearly independent of the size of the molecule. The prediction quality mainly depends on how well the query structure is represented by the molecules of the training set and of the experimental quality of the training spectra. For typical mainstream structures, the results are very convincing.

The TeleSpec project combines this novel spectrum prediction method with the methodologies of the Internet to supply a rapid and direct access channel to reference spectra. From any network-connected computer, regardless whether PC or workstation, interactive spectrum prediction experiments can be performed, and our algorithms can be tested with practical examples. We consider this way of presenting our research results to a broader audience as a great opportunity. It has never existed before in the history of the dissemination of scientific knowledge.

I wish to thank *W.-D. Ihlenfeldt* for extensive editorial help in the process of writing this paper.

Received: September 11, 1998

[1] J.H. Schuur, P. Selzer, J. Gasteiger, 'The Coding of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activitiy', *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.

[2] J. Gasteiger, J. Zupan, 'Neuronale Netze in der Chemie', *Angew. Chem.* **1993**, *105*, 510; *ibid., Int. Ed. Engl.* **1993**, *32*, 503.

[3] J. Sadowski, J, Gasteiger, G. Klebe, 'Comparison of Automatic Three-Dimensional Model-Builders Using 639 X-Ray Structures', *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000.

[4] J. Gasteiger, M. Marsili, 'Iterative Partial Equalisation of Orbital Electronegativity – A Rapid Access to Atomic Charges', *J. Chem. Soc., Perkin Trans. 2* **1984**, 559.

[5] M.D. Guillen, J. Gasteiger, 'Extension of the Method of Iterative Partial Equalization of Orbital Electronegativity to Small Ring Systems', *Tetrahedron* **1983**, *39*, 1331.

[6] The experimental data were taken from the SpecInfo IR database.

[7] D. Weininger, 'SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules', *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31. SMILES pages in the Internet: http://www.daylight.com/dayhtml/smiles/smiles-intro.html; http://www2.ccc.uni-erlangen.de/services/smiles.html.

[8] Developed by Dr. *Peter Ertl, Novartis Crop Protection AG*, CH–4002 Basel.

[9] A. Krassavine, *Chimia* **1998**, *52*, 668.

[10] JAVA Applet, courtesy of *Cherwell Scientific.*

[11] R.S. McDonal, P.A. Wilks, 'CAMP-DX: A Standard Form for the Exchange of Infrared Spectra in Computer-Readable Form', *J. Appl. Spectrosc.* **1988**, *42*, 151.