FULL PAPER

A Combined Application of Reaction Prediction and Infrared Spectra Simulation for the Identification of Degradation Products of *s*-Triazine Herbicides

Thomas Kostka, Paul Selzer, and Johann Gasteiger*[a]

Abstract: Substance identification in analytical chemistry is usually performed by comparing an experimental spectrum with a reference spectrum. Especially in environmental chemistry, reference spectra from databases are only available for a limited number of compounds. The combination of the reaction prediction system EROS and of infrared spectra simulation is a powerful tool for computer-assisted substance identification. First, possible degradation products of a chemical are predicted and then the infrared spectra of all these compounds are simulated. Comparison of the simulated infrared spectra with experimental spectra allows one to identify the structure of compounds. The method is demonstrated with the example of *s*-triazine herbicides.

Keywords: computer chemistry · IR spectroscopy · reaction prediction · spectra simulation · *s*-triazines

Introduction

Society has become increasingly concerned about the potential threat stemming from chemicals introduced into the environment. Danger cannot only be exerted by the released chemicals but also by their degradation products. Thus, the identification of the degradation products of chemicals has become of primary importance.

Clearly, the prediction of the degradation of chemicals in the environment by first principles will stay an insurmountable problem for a long time. Too many factors, such as soil composition, influence of microorganisms, distribution and adsorption mechanisms, make the prediction of the degradation reactions and mechanisms a difficult task. In such a situation, a modeling of the degradation reactions has to be built on data gained on direct observations: The identification of degradation products can be utilized for the development of a model for a degradation pathway. The model can then be used for the prediction of degradation products of additional compounds (Figure 1).

Comparison of the observed products with those predicted can then lead to a refinement of the model and an increase in the prediction accuracy. Thus, product identification and product prediction are closely intertwined in this inductive learning process.
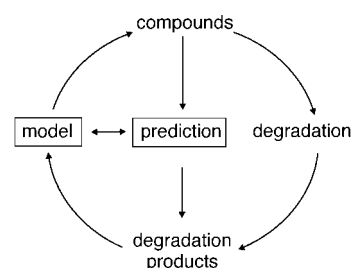

Figure 1. Learning from observations to improve the prediction power of the model.

In order to go one step further, we will show that also product identification can be facilitated and assisted by computer methods. In our case we will integrate a system for the simulation of infrared spectra into this learning process. The virtual scenario to be studied assumes that a compound is introduced into the environment and is undergoing degradation. Samples are taken from the environment and are analyzed by infrared spectroscopy after separation of the components. The analysis of this scenario should be assisted by simulation methods, by a system for the simulation of chemical reactions and by a system for the simulation of infrared spectra (Figure 2).

The reaction prediction system will give suggestions for the potential degradation products of the initial compound. For each of these compounds an infrared spectrum will be simulated. Comparison of these simulated infrared spectra with the experimental spectra obtained from the sampling procedure should then allow a rapid identification of the observed products. This, in turn, can then form the basis for a further refinement of the reaction modeling.

[a] Prof. Dr. J. Gasteiger, T. Kostka, P. Selzer
Computer-Chemie-Centrum and Institute for Organic Chemistry
University of Erlangen-Nürnberg
Nägelsbachstrasse 25, 91052 Erlangen (Germany)
Fax: (+49) 9131-85-26566
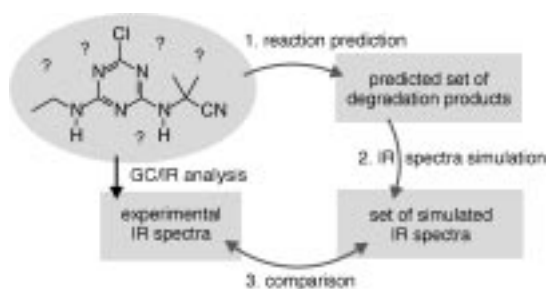E-mail: Gasteiger@chemie.uni-erlangen.de

Figure 2. The approach: combining reaction prediction and the simulation of infrared spectra to identify metabolites. Reaction prediction generates a set of possible degradation products. For each of these compounds an infrared spectrum is then simulated. Comparison of the excperimental spectrum with each of the simulated spectra allows one to identify metabolites.

We will illustrate such an attempt at the prediction of degradation reactions and products with *s*-triazine compounds, an important class of herbicides used in large amounts all over the world.[1]

Methods

Reaction prediction: The program system EROS (elaboration of reactions for organic synthesis), developed over the last 20 years, is capable of generating a wide variety of reaction types observed in organic chemistry.[2] So-called "reaction rules" store the chemical knowledge external from the program's core system. Any chemical reaction can be described as the breaking of one or more bonds followed by the formation of one or more new bonds. The computer representations of this knowledge on chemical reaction schemes forms the knowledge base of the system (Figure 3).
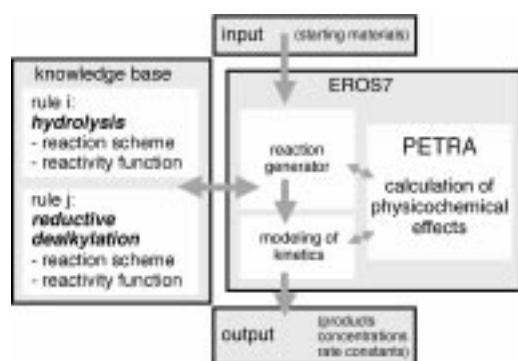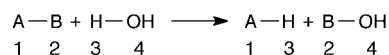

Figure 3. Functionality of EROS7—An overview: From a set of starting materials EROS generates reactions and products based on the chemical knowledge incorporated in the knowledge base.

This knowledge on chemical reactions is kept separate from the EROS system in a rule file. This allows easy change and extension of reaction types known to EROS without having to change the EROS system proper. Such a rule file keeps information on the general set-up of the reaction conditions such as temperature and solvent and a more specific part for each reaction type where the rearrangement of bonds in the course of a reaction is described and a model for the prediction of reactivity can be contained.

Figure 4 shows the general reaction scheme for hydrolysis. The breaking of the bonds between atoms 1 and 2 as well as 3 and 4 is followed by the formation of new bonds between

Figure 4. Formal reaction schemes for hydrolysis and reductive dealkylation, the main degradation pathways of 1,3,5-triazines in soil. The breaking of the bonds between atoms 1 and 2 as well as 3 and 4 is followed by the formation of new bonds between atoms 1 and 3 and atoms 2 and 4.

atoms 1 and 3 and atoms 2 and 4. Constraints concerning the types of atoms or physicochemical properties of the atoms and bonds at the reaction center can be given to limit the scope of such a reaction scheme. The reaction scheme for reductive dealkylation, an important degradation reaction of chemicals in the environment, can be specified in an analogous manner. Whenever a substructure described in the reaction rule is found in the ensemble of starting materials all conceivable corresponding reaction products will be generated.

The current program version EROS7 uses a new data structure for handling molecules and reactions which goes beyond a connection table (valence bond) representation of molecules.[3] The new data structure, MOSES, (molecular orbital structure encoding system) is based on electron systems comprising any number of electrons and atoms and therefore able to handle not only classical VB systems but also species with delocalized π systems or multi-center bonds, such as boron compounds or organometallic complexes.

Whenever sufficient experimental reactivity data is available to construct a correlation between the chemicals' properties and their reactivity, such a model can be incorporated into the rule file. This can be a mathematical function or a neural network trained with reactivity data. Such evaluation schemes can then be accessed by the reaction generator to control the number of reactions generated. Physicochemical properties of atoms and bonds (e.g., polarizability, electronegativity, σ-, π- or total charges) can be used in reactivity models and can be promptly calculated for each reaction with the PETRA (parameter estimation for the treatment of reactivity applications)[4–6] subsystem. Whenever a reactivity model has been developed and incorporated in a rule for a reaction type, EROS7 will not only generate all possible reaction products but also calculate product concentrations and rate constants. Each reaction product can act for its part as a starting material in the next reaction level and, thus, EROS7 will generate a tree of reaction steps from the input starting materials.

Simulation of infrared spectra: Infrared spectra can be calculated by quantum mechanical methods, but to obtain good correspondence with experimental data fairly extensive ab initio calculations with large basis sets or a density

functional theory calculation have to be done. On the other hand, empirical correlations between substructural units and an infrared spectrum can only explain a small part of a spectrum.

We have recently shown that a novel description of the three-dimensional structure of molecules such as the 3D-MORSE code[7, 8] or through radial distribution functions (RDF)[9, 10] allows one to obtain excellent correlations with infrared spectra. As the approach has been described elsewhere[10] we only repeat the crucial steps; Equation (1) shows a radial distribution function, *g(r)*.

$$g(r) = \sum_{j=i+1}^{n} \sum_{i=1+1}^{n-1} a_i a_j \exp(-b(r-r_{ij})^2) \qquad (1)$$

In this equation, a_i and a_j are properties of the atoms, r_{ij} is the distance between these two atoms, r is a distance variable and n gives the number of atoms in the molecule. By calculating the radial distribution function, $g(r)$, only at discrete values of r, a fixed number of descriptors of the three-dimensional structure of a molecule can be obtained. The more values of r are chosen, the higher the resolution of the three-dimensional structure representation. The discrete form of a radial distribution function will be called RDF code. In order to obtain a description of a molecule by a fixed number of descriptors, the RDF is made discrete, computing it at a fixed number of equidistant values of r.

The relationship between this RDF code and an infrared spectrum is established by a counterpropagation (CPG) neural network.[11] Figure 5 shows the architecture of a CPG network.

The top, input block will consist of structures as represented by the RDF code [see Eq. (1)]; in this application the RDF code was calculated at 128 equidistant values between 0 and 12.8 Å. The bottom, output block consists of 128 absorbance values taken from the infrared spectrum in the range of 3500 – 552 cm^{-1}. The simulated infrared spectrum shown in Figure 5 was obtained from a CPG network of size 10×10 neurons, trained with 48 structure – infrared spectra pairs taken from the SpecInfo database. The query structure was taken from a test set and it was not included in the training set.

Results

Applications to the degradation of *s*-triazine herbicides

The combined application of a system for reaction simulation and a system for infrared spectra simulation will be illustrated with two *s*-triazine herbicides. This example will show the major steps that have to be performed in order to arrive at a good predictive system. First, for a given type of compound the main reaction types for degradation have to be identified. Reactivity models should then be developed when sufficient experimental data are available in the literature. We will first perform the simulation of reactions occurring during degradation without a reactivity model. Then, we will introduce a reactivity model in order to show the enhanced performance of such a combined system when more information is introduced.

Reaction simulation without a reactivity model: Two major reaction types are involved in the degradation of *s*-triazine herbicides in soil: reductive dealkylation and hydrolysis (see Figure 4). These two reaction types were added as knowledge base to the EROS system.

Input of cyanazine to the EROS system thus conditioned produced the tree of reaction products shown in Figure 6.

As no quantitative evaluation of reactivity is included in the rules for hydrolysis and reductive dealkylation, all of those reaction pathways are likely to occur and all of those compounds may be observed.

For three of the compounds included in this scheme of reaction steps infrared spectra were found in the spectra database SpecInfo.[12] Let us assume that these spectra would have been available from an experiment in which these compounds had been isolated from the environment. The question is then: Would it have been possible to identify the corresponding compounds among this set of 28 rather similar compounds solely on the basis of their infrared spectra?

In order to explore this question, an infrared spectrum was simulated for each of the compounds depicted in Figure 6. To this effect all *s*-triazine compounds availabe from the SpecInfo database were collected. Three of these compounds were identified with compounds also found in Figure 6 and were therefore removed from this set.

The remaining 46 compounds of the dataset were used together with their corresponding infrared spectra to train a counterpropagation (CPG) network of 10×10 neurons. Then, each of the structures depicted in Figure 6 was sent through this CPG network, thus providing a simulated infrared spectrum for each of the 28 compounds of Figure 6.

Each of the three experimental spectra was then compared with every one of the 28 simu-
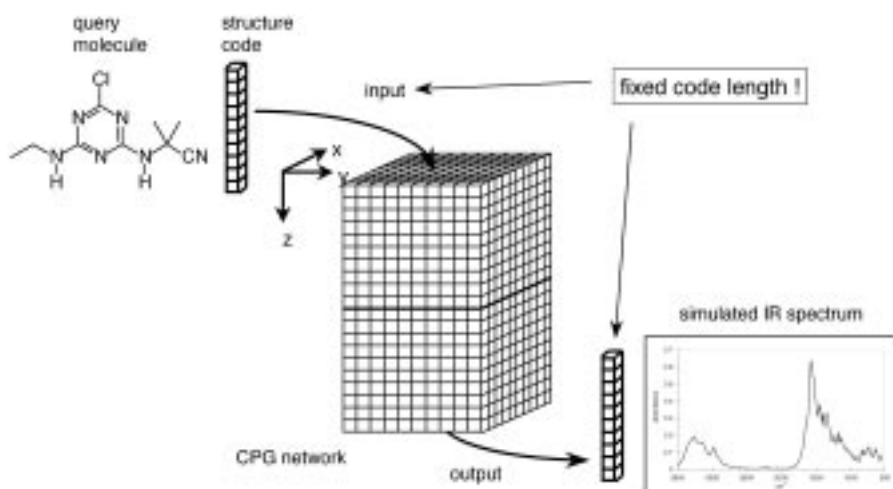
Figure 5. Simulation of an infrared spectrum by a CPG neural network. The structure code of the query molecule is used to retrieve a simulated spectrum from the trained network.
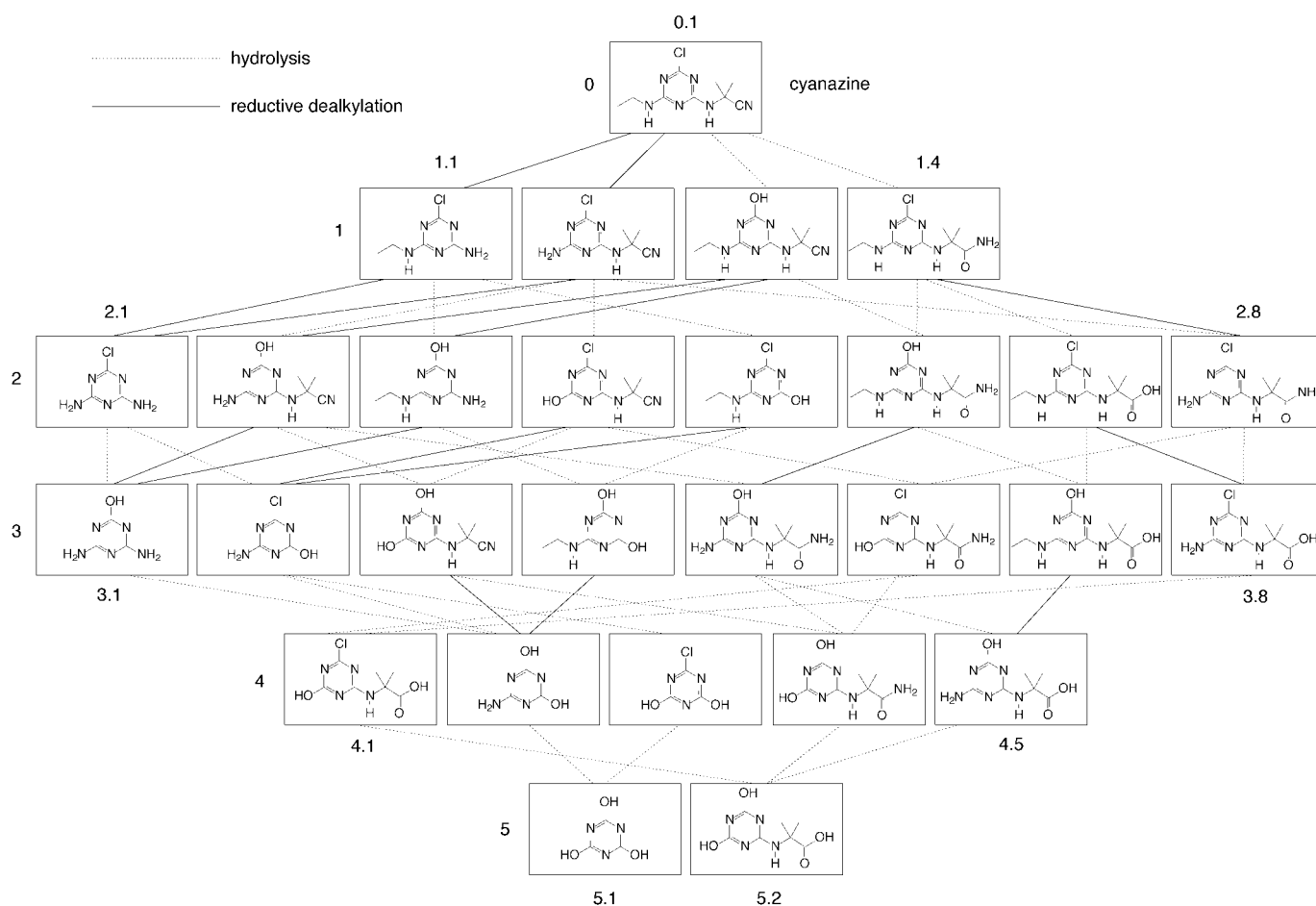
Figure 6. Tree of simulated degradation reactions for cyanazine with hydrolysis and reductive dealkylation as main degradation pathways in soil. Five levels of degradation reactions are shown.

lated spectra to determine which simulated infrared spectrum had highest correspondence with the selected experimental spectrum. The Pearson correlation coefficient[13] was taken as a measure of similarity. Figure 7 shows the results thus obtained. The bold frames identify the simulated spectra with closest resemblence to the respective experimental spectrum.

The question is whether this procedure gives correct structural assignments. Figure 8 shows the results that were obtained by this kind of comparison of the experimental with simulated infrared spectra.

In the first two cases the proposed structures were correct. In the third case, the proposed structure contained two amino and one hydroxy group on the triazine system, whereas the experimental spectrum was actually from the compound having one amino and two hydroxy groups. The low correlation coefficient r of 0.719 indicated that the assigned structure is less reliable. The reason for this lower correspondence is a lack of information for the training process since the database contained no molecule with an OH-substituted s-triazine.

As a next example, the same procedure was performed with trietazine. Figure 9 shows the potential degradation products obtained by application of the two reaction types, reductive dealkylation, and hydrolysis, as in the previous investigation. The infrared spectra were simulated for all compounds as outlined above. All structures from the SpecInfo database

which were found to be identical with one of the structures depicted in Figure 9, in this case again three compounds were deleted from the training set and used for testing the quality of the structure prediction by comparing their spectra with each one of the 18 simulated spectra.

The experimental spectra of the three compounds of Figure 9 which were in the database were compared in turn with each one of the 18 simulated spectra. Figure 10 ranks the simulated spectra according to their correlation coefficients with each one of the experimental spectra.

The correct structures in this example are at position 2, 3, and 3, respectively, or, if tautomers are also taken into account, the correct structures take the ranks 2, 3, and 2, respectively. Can this situation be improved and can the correct structures eventually obtain a higher ranking? We will show in the next section that a more detailed modeling of the reaction types which also incorporates a quantitative evaluation of chemical reactivity can improve the prediction accuracy.

Reaction simulation with a reactivity model: The development of a reactivity model needs observations and data on chemical reactions. The more detailed and accurate data are available, the better a model can be developed. The ideal case would be if detailed kinetic measurements on a series of

experimental
spectra

a)
r = 0.885

b)
r = 0.987

c)
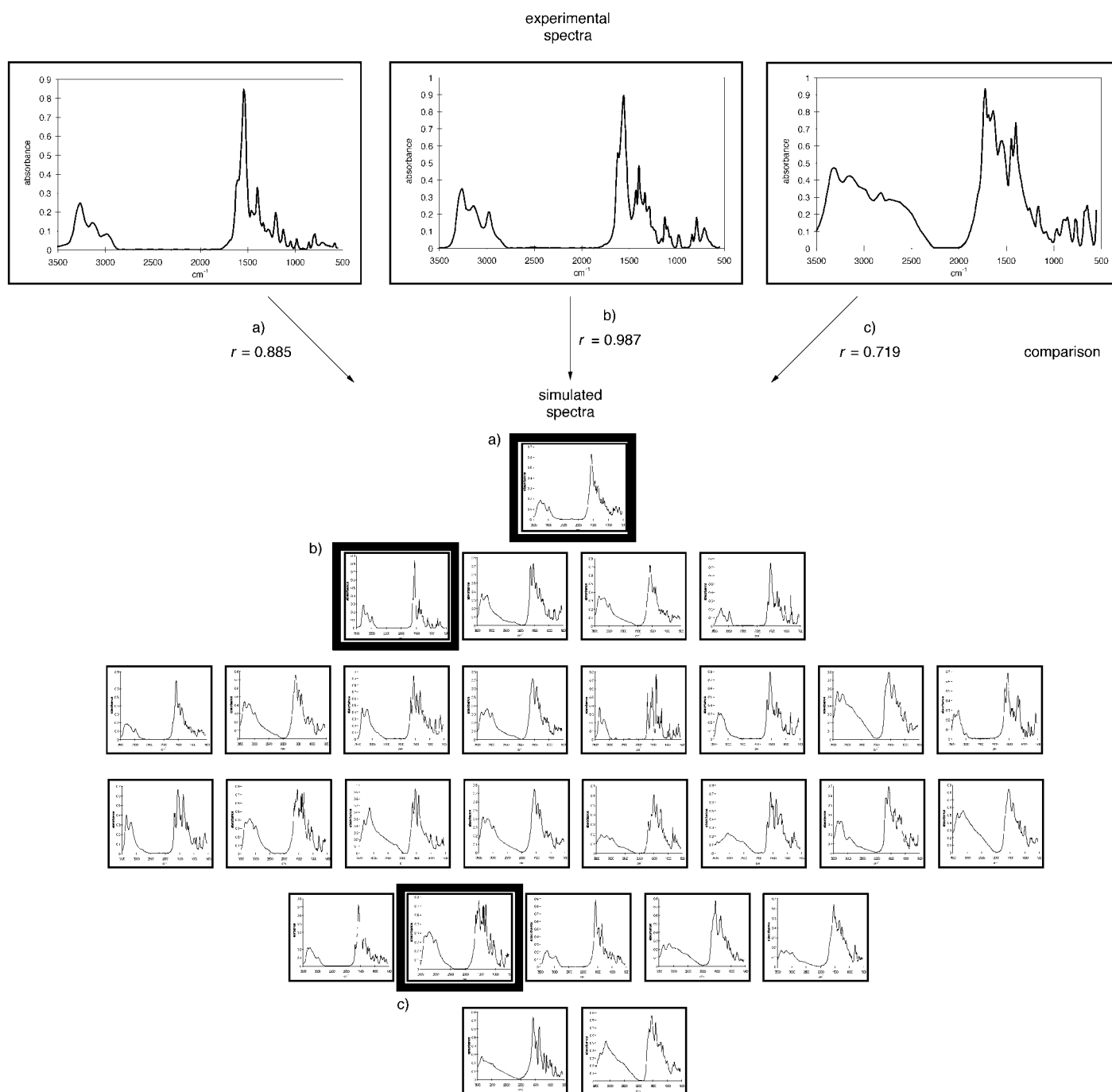r = 0.719   comparison

simulated
spectra

Figure 7. Comparison of experimental spectra with the simulated IR spectra for all cyanazine metabolites shown in Figure 6. Bold frames indicate best correlation between experimental and simulated spectra.

reactions had been performed. Such data, however, could not be found for the degradation of *s*-triazine compounds similar to those investigated here. The best data we could find were half-life times of four *s*-triazine herbicides for their degradation in soil (see Figure 11).[14]

For the reactivity model two assumptions were made. First, the rate of hydrolysis was considered as constant and independent of the structural differences in the amino-alkyl groups. Secondly, both reaction types, hydrolysis and reductive dealkylation, should follow first order kinetics. Each overall degradation rate will then be the sum of the degradation rate for hydrolysis and the rates of dealkylation for the two alkyl groups. A constant rate of hydrolysis of $k = 5.10^{-9}$ s^{-1} was assumed, based on hydrolysis rates reported in the literature[15] and adapted to the scenario of our four herbicides.[14] With the rate of hydrolysis known, the reaction rate for the dealkylation of an ethyl group can be calculated from the half-life of simazine. Analogously, the rate constants for the dealkylation of an isopropyl and *tert*-butyl group can be derived from the half-life of propazine and terbuthylazine.
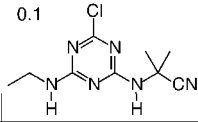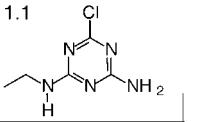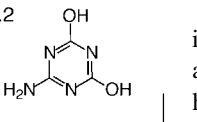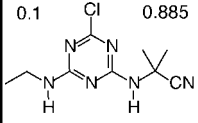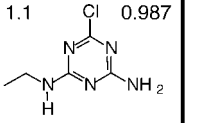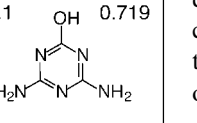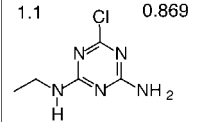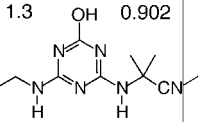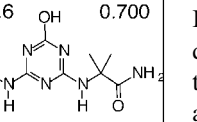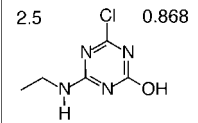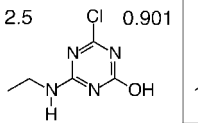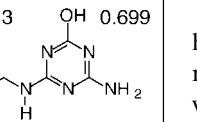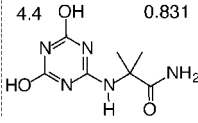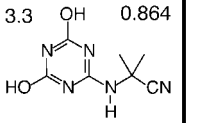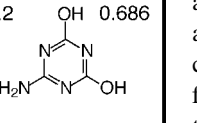
Figure 8. Ranking of the correct structures using Pearson correlation coefficient as shown at the upper right-hand side of each structure. Bold frames indicate the correct structures. Top row: Structures for experimentally observed compounds. Ranks indicate simulated spectra.

The information on the individual reaction rates was then implemented into the EROS rule file. We already used this approach of modeling degradation kinetics for triazine herbicides in combination with exposure modeling for the development of a software system for risk assessment of chemicals in soil.[16] Figure 12 shows the calculated concentrations of each compound from Figure 9 during a period of one year.

One can clearly see that the concentration of compound III.3 is rather low during this one-year period, so that we can delete compound III.3 in our ranking of Figure 10. Therefore, the ranking of the correct structures improves to ranks 1, 2, and 2.

It should be pointed out that the structures investigated here are quite similar and often have infrared spectra that are rather close. For a human expert the task presented here would be very difficult if not impossible to solve. Clearly, this approach depends on the amount and quality of data available. The database used as a basis for the investigation contained 49 s-triazine derivatives. Only four infrared spectra for the overall 36 structures investigated were available from the database. Thus, quite accurate intrapolations were necessary to obtain the high quality simulated infrared spectra needed here. Efforts to collect infrared spectra over the internet to improve the simulation power of this system are in progress.[17]
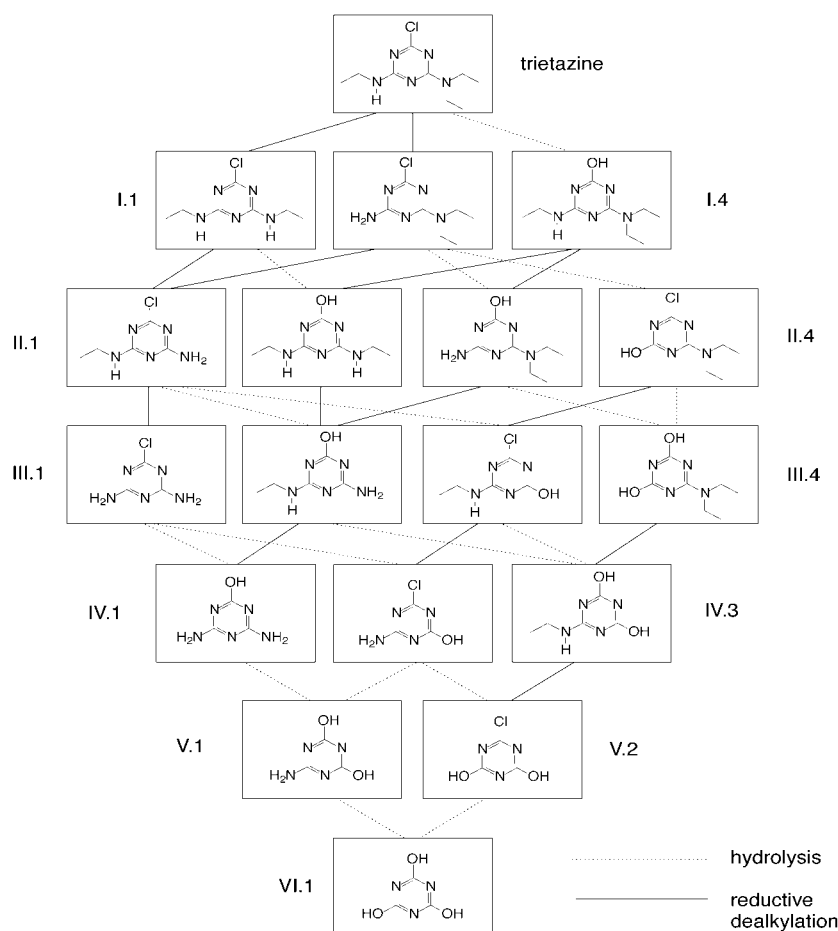

Figure 9. Tree of simulated degradation reactions for trietazine with hydrolysis and reductive dealkylation as main degradation pathways in soil. Six levels of degradation reactions are shown.
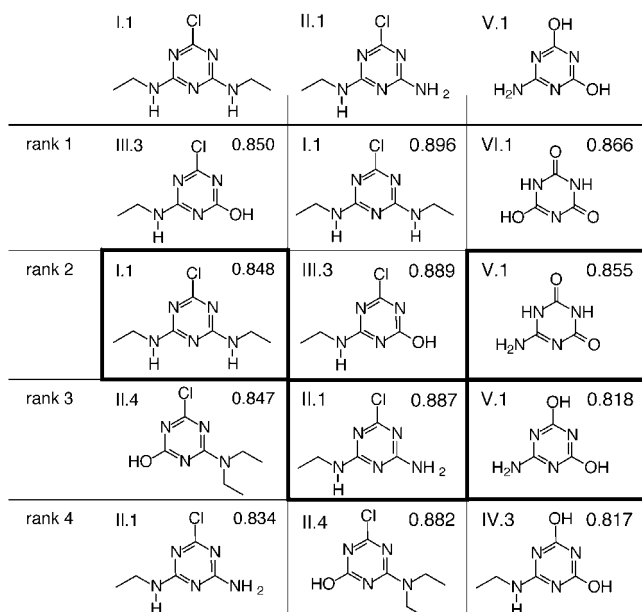
Figure 10. Ranking of simulated IR spectra for all trietazine metabolites shown in Figure 9; the Pearson correlation coefficient is shown at the upper right-hand side of each structure. Bold frames indicate the correct structures. Top row: Structures for experimentally observed compounds. Ranks indicate simulated data.
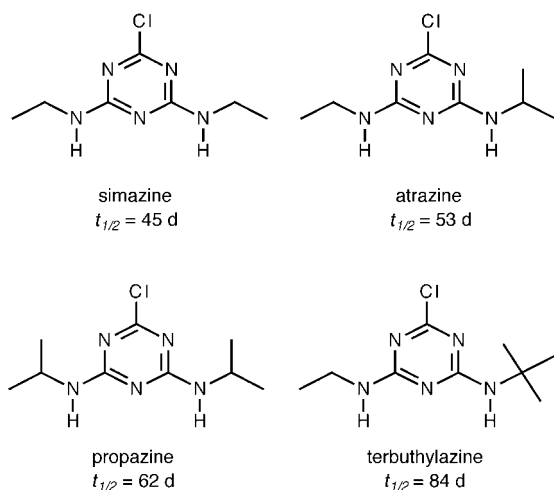

simazine
$t_{1/2} = 45$ d

atrazine
$t_{1/2} = 53$ d

propazine
$t_{1/2} = 62$ d

terbuthylazine
$t_{1/2} = 84$ d

Figure 11. Half lifes for simazine, atrazine, propazine, and terbuthylazine from ref. [14].
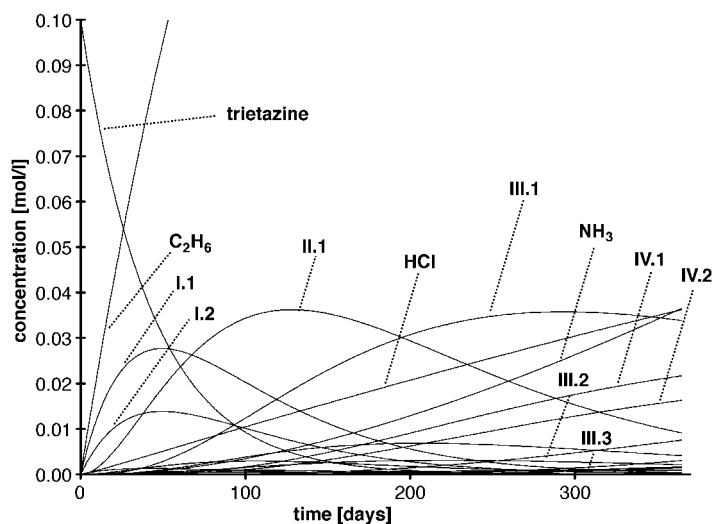

Figure 12. Concentration versus time plots for trietazine degradation products.

knowledge on the reactivity of a class of compounds can somehow compensate for deficiencies in the database of infrared spectra and vice versa.

Conclusion

The two examples of studies on the degradation of *s*-triazine herbicides show the power of this approach both for structure elucidation and for the modeling of reaction pathways. These two tasks are closely interrelated: Product identification is the first step in modeling a reaction pathway. Knowledge on a reaction pathway on the other hand facilitates product identification.

Comparison of the results for trietazine with those for cyanazine shows that a deeper knowledge on chemical reactivity allows one to make better inferences and obtain correct structure assignments more rapidly. Thus, more

[1] Environmental Protection Agency, "The Triazine Pesticides", to be found under http://www.epa.gov/pesticides/citizens/triazine.htm, **1999**.

[2] R. Höllering, J. Gasteiger, L. Steinhauer, K.-P. Schulz, A. Herwig, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482–494.

[3] S. Bauerschmidt, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 705–714.

[4] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.

[5] M. G. Hutchings, J. Gasteiger, *Tetrahedron Lett.* **1983**, *24*, 2537–2540.

[6] Access to the PETRA program package through the Internet: http://www2.ccc.uni-erlangen.de/software/petra/

[7] J. H. Schuur, P. Selzer, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334–344.

[8] J. Schuur, J. Gasteiger, *Anal. Chem.* **1997**, *69*, 2398–2405.

[9] M. C. Hemmer, V. Steinhauer, J. Gasteiger, *Vibrat. Spectrosc.* **1999**, *19*, 151–164.

[10] P. Selzer, J. Gasteiger, H. Thomas, R. Salzer, *Chem. Eur. J.* **2000**, *6*, 920–927.

[11] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, 2nd ed., Wiley-VCH, Weinheim, **1999**.

[12] *SpecInfo Infrared Database*, Chemical Concepts, Weinheim, Germany.

[13] G. E. P. Box, W. G. Hunter, J. S. Hunter, *Statistics for Experimenters*, Wiley, New York, **1978**.

[14] N. Burckhard, J. A. Guth, *Pestic. Sci.* **1981**, *12*, 45–52.

[15] Environmental Research Lab, Athens, Ga, Publication No. EPA-600/3-76-067, **1976**, 125–128.

[16] H. Behrendt, J. Altschuh, S. Sixt, J. Gasteiger, R. Höllering, T. Kostka, *Chemosphere* **1999**, *38*, 1811–1823.

[17] IR spectrum simulation via the internet—TeleSpec: http://www2.ccc.uni-erlangen.de/IR/