Linking Databases of Chemical Reactions to NMR Data: an Exploration of 1H NMR−Based Reaction Classification

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2007

Impact Factor: 5.64 · DOI: 10.1021/ac060979s

2 AUTHORS:

Diogo ARS Latino
Eawag: Das Wasserforschungs-Institut des E…
41 PUBLICATIONS **150** CITATIONS
SEE PROFILE

João Aires-de-Sousa
New University of Lisbon
71 PUBLICATIONS **1,043** CITATIONS
SEE PROFILE

Anal. Chem. **2007,** *79,* 854−862

Linking Databases of Chemical Reactions to NMR Data: an Exploration of ^{1}H NMR-Based Reaction Classification

Diogo A. R. S. Latino and João Aires-de-Sousa*

CQFB and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Automatic analysis of changes in the ^{1}H NMR spectrum of a mixture and their interpretation in terms of chemical reactions taking place have a diversity of possible applications, from the monitoring of reaction processes or degradation of chemicals to metabonomics. Classification of photochemical and metabolic reactions by Kohonen self-organizing maps and random forests is demonstrated, taking as input the difference between the ^{1}H NMR spectra of the products and the reactants. The chemical shifts of the reactants and products were fuzzified to obtain a crude representation of the spectra. With a dataset of 911 metabolic reactions catalyzed by transferases (EC number 2.x.x.x), classification according to subclass (second digit of the EC number) could be achieved with up to 84% of accuracy. Experiments with a dataset of 189 photochemical reactions, manually assigned to seven classes, yielded 86−93% of correct classifications for an independent test set of 42 reactions, and the models were further validated with a test set combining experimental and simulated chemical shifts. The results support our proposal of linking databases of chemical reactions to NMR data for automatic reaction classification and show the usefulness of the predictions obtained by the SPINUS program for the estimation of missing NMR experimental data.

Processing ^{1}H NMR data by machine learning techniques is expanding the application of this spectroscopy to domains far beyond the classical use in structure elucidation.[1] Most notably, ^{1}H NMR spectroscopy is playing a central role in the emerging area of metabonomics,[2] in which complex multivariate data from biological samples are analyzed by machine learning and statistical methods, in order to detect certain metabolites and potential biomarkers or to assess disease pathology, drug efficacy, toxicological profiles, gene expression, and mode of action of bioactive compounds.[3−9] Lindon[10] reviewed applications of pattern recognition methods in biomedical magnetic resonance. These include unsupervised learning methods, such as principal component analysis, nonlinear mapping, and hierarchical cluster analysis, for reduction of data complexity, and supervised learning methods for sample classification, such as SIMCA, partial least-squares, linear discriminant analysis, and artificial neural networks.

Kohonen self-organizing maps (SOMs)[11,12] have been used to process ^{1}H NMR data in varied applications. Beckonert et al.[13] used ^{1}H NMR spectroscopy and SOMs to study metabolic changes in breast cancer tissue. Van et al.[14] distinguished individuals affected by interstitial cystitis and by bacterial cystitis from nonaffected individuals on the basis of mass spectrometry and ^{1}H NMR spectral patterns of urine. Bathen et al.[15] investigated SOMs for the classification of ^{1}H NMR spectra of the human blood plasma lipoprotein fractions from healthy volunteers and patients with cancer or coronary heart disease.

The usefulness of ^{1}H NMR spectroscopy in reaction and process monitoring has also been demonstrated; for example, in liquid-phase combinatorial synthesis,[16] estimation of velocities of enzyme-catalyzed reactions,[17] or using on-line NMR spectros-

* Corresponding author. Phone: (+351) 21 2948300. Fax: (+351) 21 2948550. E-mail: jas@fct.unl.pt.

(1) Alam, T. M.; Alam, M. K. Chemometric Analysis of NMR Spectroscopy Data: A Review. In *Annual Reports on NMR Spectroscopy*; Webb, G.A., Ed.; Academic Press: London, 2004; Vol. 54, pp 41−80.
(2) Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *NMR Biomed.* **2005**, *18*, 143−162.
(3) Espina, J. R.; Shockcor, J. P.; Herron, W. J.; Car, B. D.; Contel, N. R.; Ciaccio, P. J.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Magn. Reson. Chem.* **2001**, *39*, 559−565.
(4) Holmes, E.; Nicholls, A.; Lindon, J. C.; Connor, S. C.; Connelly, J. C.; Haselden, J. N.; Damment, S. J. P.; Spraul, M.; Neidig, P.; Nicholson, J. K. *Chem. Res. Toxicol.* **2000**, *13*, 471−478.
(5) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181−1189.
(6) Holmes, E.; Nicholls, A. W.; Lindon, J. C.; Ramos, S.; Spraul, M.; Neidig, P.; Connor, S. C.; Connelly, J.; Damment, S. J. P.; Haselden, J.; Nicholson, J. K. *NMR Biomed.* **1998**, *11,* 235−244.
(7) Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Everett, J. R. *Concepts Magn. Reson.* **2000**, *12*, 289−320.
(8) Holmes, E.; Nicholson, J. K.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Polley, S.; Connelly, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 245−255.
(9) Ott, K.; Aranibar, N.; Singh, B.; Stockton, G. *Phytochemistry* **2003**, *62*, 971−975.
(10) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Prog. NMR Spectrosc.* **2001**, *39*, 1−40.
(11) Kohonen, T. *Self-Organization and Memory*; Springer: Berlin, 1988.
(12) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, 1999.
(13) Beckonert, O.; Monnerjah, J.; Bonk, U.; Leibfritz, D. *NMR Biomed.* **2003**, *16*, 1−11.
(14) Van, Q. N.; Klose, J. R.; Lucas, D. A.; Prieto, D. A.; Luke, B.; Collins, J.; Burt, S. K.; Chmurny, G. N.; Issaq, H. J.; Conrads, T. P.; Veenstra, T. D.; Keay, S. K. *Dis. Markers* **2003**, *19*,169−183.
(15) Bathen, T. F.; Engan, T.; Krane, J.; Axelson, D. *Anticancer Res.* **2000**, *20,* 2393−2408.
(16) Shey, J.-Y.; Sun, C.-M. *Tetrahedron Lett.* **2002**, *43*, 1725−1729.

10.1021/ac060979s CCC: $37.00
Published on Web 12/20/2006

copy.[18] Kalelkar et al. described a SOM analysis of ^{1}H NMR spectra from combinatorial parallel synthesis with the aim of identifying outliers in the libraries.[19] However, to the best of our knowledge, SOMs have not been investigated for reaction classification on the basis of NMR data.

The SPINUS program[20−22] for the estimation of ^{1}H NMR chemical shifts from the molecular structure allows linking a database of chemical reactions to the corresponding ^{1}H NMR data. In this paper, we explore the classification of chemical reactions by Kohonen self-organizing maps and random forests (RFs),[23] taking as input the difference between the ^{1}H NMR spectra of the products and the reactants. The rationale behind this proposal is that the substructures of the reactants that are far from the reaction center mostly will have their chemical shifts unchanged, whereas the chemical shifts of the atoms near the reaction center should change with the reaction. The pattern of changes can be interpreted as a descriptor of the reaction. Such a representation additionally has the potential to encode 3D effects, even if related to substructures topologically distant from the reaction center, and can, in principle, be applied when more than one reaction occurs simultaneously. Clearly, ^{1}H NMR spectroscopy has its own limitations, particularly for reactions with a small number of hydrogen atoms in the neighborhood of the reaction center. At the same time, the use of ^{1}H NMR has considerable advantages in comparison to NMR of other nuclei, such as the speed and the amount of sample required.

Automatic analysis of changes in the ^{1}H NMR spectrum of a mixture and their interpretation in terms of chemical reactions taking place have a diversity of possible applications. The changes in the ^{1}H NMR spectrum of a stored chemical can be interpreted in terms of the chemical reactions responsible for degradation. A database of metabolic reactions linked to the corresponding NMR data can assist in the monitoring of a biotechnological process by ^{1}H NMR spectroscopy and yield information about the enzymatic reactions taking place, or the alterations in the spectrum of a biofluid can be related to changes in metabolic reactions.

In a related but conceptually different direction, the automatic classification of chemical reactions from the molecular structures of the reactants and products is currently a topic of high interest in chemoinformatics, particularly in relation to genome-scale study of metabolic reactions and pathways.[24,25] Several methods for the representation and classification of reactions have been put forward,[26−39] mostly based on codes of the reaction center or on physicochemical properties of atoms and bonds at the reaction center. Most of these methods require atom-to-atom mapping and assignment of the atoms or bonds involved in the reaction. In our lab, a strategy has been devised for the representation of chemical reactions without assignment of the reaction centers[40] that is based on differences between molecular descriptors of the products and the reactants. Recently, this approach could be applied to genome-scale classification of metabolic reactions.[41]

In the work here described, we explore the possibility of inferring the type of reaction exclusively from ^{1}H NMR spectra. To address this problem, we simulated a situation in which the structures of the reactants and products are unknown, but their ^{1}H NMR spectra are available. Two machine learning techniques, Kohonen SOMs and random forests, were investigated to classify photochemical and metabolic reactions from the ^{1}H NMR chemical shifts of the products and the reactants. These machine learning methods differ in the type of learning. Whereas Kohonen SOMs are trained with unsupervised learning (competitive learning), random forests are trained with supervised learning.

We first explored a dataset of photochemical cycloadditions. This was partitioned into a training set of 147 reactions and a test set of 42 reactions, all manually classified into seven classes. The ^{1}H NMR spectra were simulated from the molecular structures by SPINUS,[20−22] and a reaction was represented by the difference between the spectrum of the product and the spectra of the reactants. After the predictive models for the classification of chemical reactions were established on the basis of simulated NMR data, their applicability to reaction data from mixed sources (experimental and simulated) was evaluated.

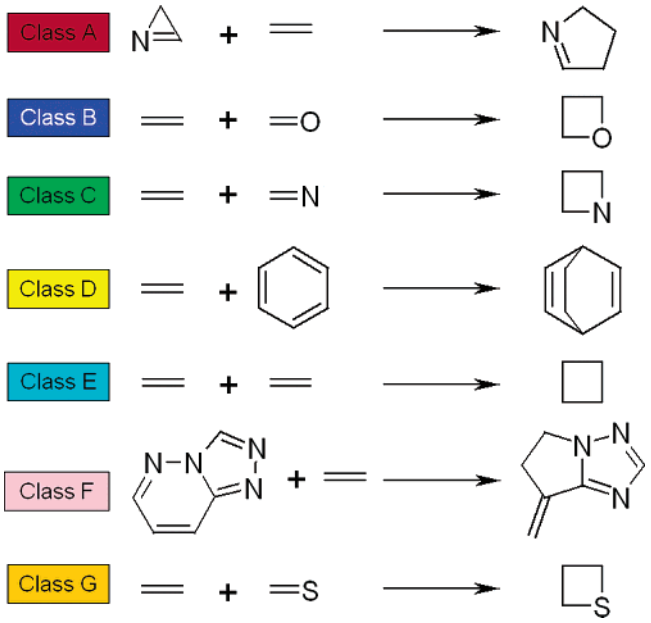
A second dataset was also explored, consisting of 911 metabolic reactions catalyzed by transferases (EC number 2.x.x.x) classified into eight subclasses according to the Enzyme Commission (E.C.) system.[42] Differently from the first dataset, not all the reactions have the same number of reactants and products. Apart from that, automatic classification was investigated in the same way.

METHODOLOGY

The experiments here described involve two steps: the generation of a reaction descriptor from the simulated ^{1}H NMR

(17) Vallikivi, I.; Jarving, I.; Pehk, T.; Samel, N.; Tougu, V.; Parve, O. *J. Mol. Catal. B: Enzym.* **2004**, *32*,15−19.
(18) Maiwald, M.; Fischer, H. H.; Kim, Y. K.; Albert, K.; Hasse, H. *J. Magn. Reson.* **2004**, *166*, 135−146.
(19) Kalelkar, S; Dow, E. R.; Grimes, J.; Clapham, M.; Hu, H. *J. Comb. Chem.* **2002**, *4*, 622−629.
(20) Binev, Y.; Aires-de-Sousa, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 940−945.
(21) Binev, Y.; Corvo, M.; Aires-de-Sousa, J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 946−949.
(22) SPINUS can be accessed at http://www.dq.fct.unl.pt/spinus or http://www2.chemie.uni-erlangen.de/services/spinus (accessed Jan 2006).
(23) Breiman, L. *Machine Learn.* **2001**, *45*, 5−32.
(24) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. *J. Am. Chem. Soc.* **2004**, *126*, 16487−16498.
(25) Gasteiger, J.; Reitz, M.; Sacher, O. *The Chemical Theatre of Biological Systems*; Proceedings of the Beilstein-Institut Symposium, Bozen, Italy, 2004.
(26) Chen, L. Reaction Classification and Knowledge Acquisition. In *Handbook of Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: New York, 2003; Vol. 1, pp 348−388.
(27) Rose, J. R.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74−90.
(28) Chen, L.; Gasteiger, J.; Rose, J. R. *J. Org. Chem.* **1995**, *60*, 8002−8014.
(29) Chen, L.; Gasteiger, J. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.
(30) Sacher, O. Ph.D. Thesis, University of Erlangen-Nuremberg, 2001; http://www2.chemie.uni-erlangen.de/services/dissonline/data/dissertation/Oliver_Sacher/html/ (accessed Jan 2006).
(31) Fujita, S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205−212.
(32) Hendrickson, J. B. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 852−860.
(33) Tratch, S. S.; Zefirov, N. S. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 349 -366.
(34) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1998**, *38,* 210−219.
(35) *Classify—The InfoChem Classification Program V. 2.5*; to be found under http://www.infochem.de/content/downloads/classify.pdf (accessed Nov 2006).
(36) Chen, L.; Gasteiger, J. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 763−765.
(37) Satoh, H.; Itono, S.; Funatsu, K.; Takano, K.; Nakata, T. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 671 -678.
(38) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiau, J. *Tetrahedron Comput. Methodol.* **1988**, *1*, 117−128.
(39) Dugundji, J.; Ugi, I. *Top. Curr. Chem.* **1973**, *39*, 19−64.
(40) Zhang, Q.-Y.; Aires-de-Sousa, J. *J. Chem. Inf. Model.* **2005**, *45,* 1775−1783.
(41) Latino, D. A. R. S.; Aires-de-Sousa, J. *Angew. Chem., Int. Ed.* **2006**, *45*, 2066−2069.
(42) Barrett, A. J.; Canter, C. R.; Liebecq, C.; Moss, G. P.; Saenger, W.; Sharon, N.; Tipton, K. F.; Vnetianer, P.; Vliegenthart, V. F. G. *Enzyme Nomenclature*; Academic Press: San Diego, CA, 1992.

Scheme 1. Classes of Photochemical Reactions



spectra of the products and reactants and the development of predictive models for reaction classification using Kohonen self-organizing maps or random forests.

Datasets of Reactions. A dataset of 189 photochemical reactions, involving two reactants and one product (bearing at least one hydrogen atom covalently bonded to a carbon atom) was extracted from the SPRESI database (InfoChem GmbH, Munich, Germany). The reactions were manually assigned into seven classes: [3 + 2] photocycloaddition of azirines to C=C (class A, 20 reactions), [2 + 2] photocycloaddition of C=C to C=O (class B, 31 reactions), [2 + 2] photocycloaddition of C=N to C=C (class C, 8 reactions), [4 + 2] and [4 + 4] photocycloaddition of olefins to carbon-only aromatic rings (class D, 20 reactions), [2 + 2] photocycloaddition of C=C to C=C (class E, 73 reactions), [3 + 2] photocycloaddition of *s*-triazolo[4,3-*b*]pyridazine to C=C (class F, 10 reactions), and [2 + 2] photocycloaddition of C=C to C=S (class G, 27 reactions). Scheme 1 summarizes the different types of reactions. The dataset of 189 reactions were randomly partitioned into a training set of 147 reactions (16 of type A, 23 of type B, 7 of type C, 16 of type D, 56 of type E, 8 of type F, and 21 of type G) and a test set of 42 reactions (4 of type A, 8 of type B, 1 of type C, 4 of type D, 17 of type E, 2 of type F, and 6 of type G), assuring that both sets cover the whole range of reactions.

For validating the models with experimental ^{1}H NMR data, a subset of 26 reactions was assembled from the original set, with the simulated chemical shifts replaced by experimental values for some reactants or products. The chemical shifts were obtained from the SDBS database[43] or from the references associated with the reactions in the SPRESI database. In three reactions, experimental data was included for both reactants and the product; in 14 reactions, for only one reactant; in six reactions for only the product; in one reaction, for the product and one reactant; and in two reactions, for both reactants. The subset of 26 reactions (9 were from the test set and 17 from the training set) includes 3 reactions of class A, 3 of class B, 1 of class C, 4 of class D, 8 of class E, 2 of class F, and 5 of class G.

For the second application, a dataset was assembled with metabolic reactions catalyzed by transferases (EC number 2.x.x.x) extracted from the KEGG LIGAND database[44] of enzymatic reactions (release of January 2006). The selected dataset excluded reactions listed with more than one EC number or an incomplete EC number. Reactions involving a compound not accepted by SPINUS, as well as unbalanced reactions, were also excluded. Reactions differing only in stereochemical features were considered as duplicates and were included only once. The cleaning procedure was based on chemical hashed fingerprints generated by JChem package, version 3.1.7.1 (ChemAxon, Budapest, Hungary, www.chemaxon.com) with a length of 64 bytes, a maximum number of five bonds in patterns, and two bits switched on for each pattern in the structure. The dataset finally consisted of 911 reactions catalyzed by transferases (EC number 2.x.x.x) belonging to eight different subclasses (second digit of the EC number): 133 reactions transferring one-carbon groups (EC 2.1.x.x, here labeled with "class A"), 9 reactions transferring aldehyde or ketonic groups (EC 2.2.x.x, class B), 171 reactions catalyzed by acyltransferases (EC 2.3.x.x, class C), 201 reactions catalyzed by glycosyltransferases (EC 2.4.x.x, class D), 41 reactions transferring aryl or alkyl groups other than methyl (EC 2.5.x.x, class E), 75 reactions transferring nitrogenous groups (EC 2.6.x.x, class F), 259 reactions transferring phosphorus-containing groups (EC 2.7.x.x, class G), and 22 reactions transferring sulfur-containing groups (EC 2.8.x.x, class H). The dataset of 911 reactions was partitioned into training and test sets using a 18 × 18 Kohonen SOM. The SOM was trained with all reactions, and after training, one reaction was randomly selected from each occupied neuron and moved to the test set. With this procedure, a test set resulted with 262 reactions (40 of subclass 2.1.x.x, 3 of subclass 2.2.x.x, 60 of subclass 2.3.x.x, 62 of subclass 2.4.x.x, 15 of subclass 2.5.x.x, 24 of subclass 2.6.x.x, 52 of subclass 2.7.x.x, and 6 of subclass 2.8.x.x). The training set consisted of the remaining 649 reactions (93 of subclass 2.1.x.x, 6 of subclass 2.2.x.x, 111 of subclass 2.3.x.x, 139 of subclass 2.4.x.x, 26 of subclass 2.5.x.x, 51 of subclass 2.6.x.x, 207 of subclass 2.7.x.x, and 16 of subclass 2.8.x.x).

Simulation of ^{1}H NMR Spectra with SPINUS. The SPINUS program[20−22] was used for the estimation of ^{1}H NMR chemical shifts from the molecular structures. In these simulations, only hydrogen atoms bonded to carbon atoms were predicted. A crude representation of the spectrum was obtained by fuzzifying the predicted chemical shifts with a triangular function. The triangular function was used with widths 0.2, 0.3, and 0.4 ppm at each side of the chemical shift, which approximate the observed mean absolute error of SPINUS predictions (0.2−0.3 ppm).[21] No information concerning coupling constants was used.

^{1}H NMR-Based Descriptor of Chemical Reactions. All the signals, integrating proportionally to the number of protons, arising from all reactants of one reaction were taken together (a simulated spectrum of the mixture of reactants) and were subtracted from the simulated spectrum of the mixture of products. In the case of photochemical reactions, all reactions have two reactants and one product, but in the dataset of metabolic

(43) http://www.aist.go.jp/RIODB/SDBS/cgi-bin/cre_index.cgi (accessed Jan 2006).

(44) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.* **2006**, *34*, D354−357.

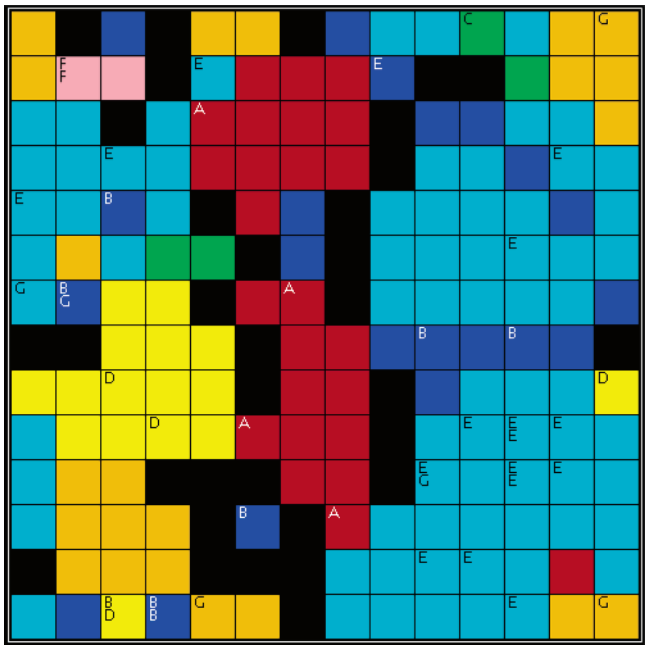


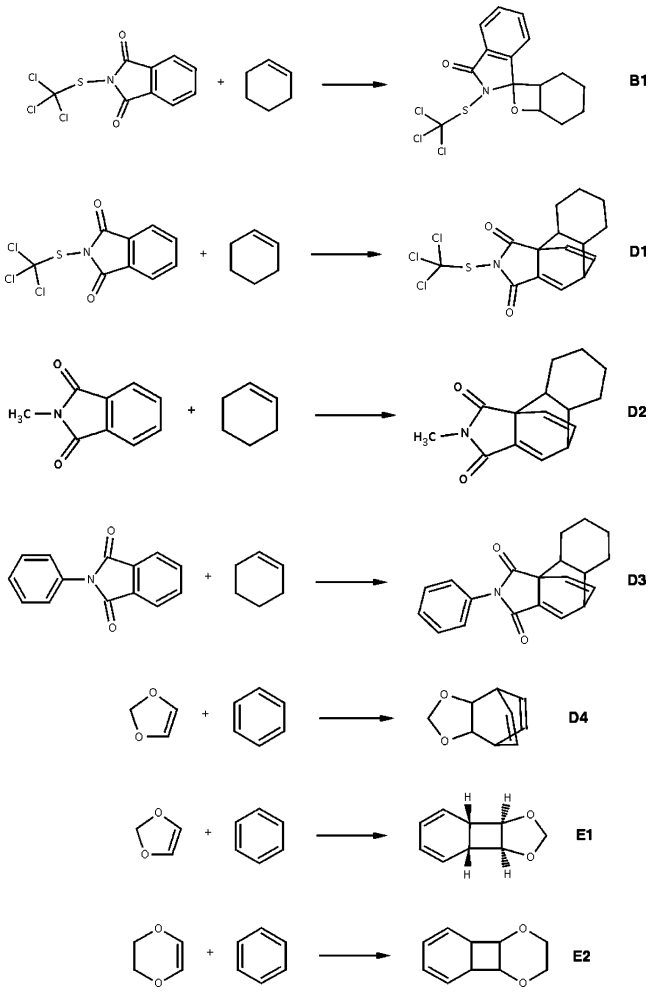
Figure 1. Toroidal surface of a 14 × 14 Kohonen self-organizing map trained with the ^{1}H NMR-based descriptor for photochemical reactions belonging to classes A−G. After the training, each neuron was colored according to the reactions in the training set that were mapped onto it or onto its neighbors. The reactions of the test set were mapped onto the trained SOM and are represented by the label of their true classes. A fuzzification parameter of 0.2 ppm was used in this experiment for the simulation of NMR spectra. Colors are defined in Scheme 1.

reactions, the number of reactants and products varies among the reactions. The difference spectrum is taken as the representation of the chemical reaction. The spectrum, covering the range of 0−12 ppm, was converted into a 120-positions code, each position integrating the intensities within an interval of 0.1 ppm.

Kohonen Self-Organizing Maps.[11,12] A Kohonen self-organizing map distributes objects over a 2D surface (a grid of neurons) in such a way that objects bearing similar features are mapped onto the same or adjacent neurons. SOMs perform a nonlinear projection of multidimensional objects onto a two-dimensional surface, yielding maps of easy visual interpretation. The input data are stored in the two-dimensional grid of neurons, each neuron containing as many elements (weights) as there are input variables. In the investigations described in this paper, the input variables are the above-mentioned 120 reaction descriptors derived from ^{1}H NMR spectra. We trained SOMs with a diversity of reactions to investigate the ability of the ^{1}H NMR descriptors to cluster and classify chemical reactions. SOMs with toroidal topology and sizes varying between 13 × 13 and 15 × 15 for photochemical reactions and 25 × 25 or 29 × 29 for metabolic reactions were trained with the training set and tested with the test set. The SOMs learn in an unsupervised way, which means that the distribution of objects on the map relies on only the NMR data, not on the preassigned classes. After the training, each neuron (a position of the map) was assigned to a class, depending on the reactions that were mapped into it, or into its neighbors if it was empty. If a winning class could not be identified for a neuron, the neuron was classified as undecided.

Training of the SOMs was performed by using a linear decreasing triangular scaling function with an initial learning rate

Scheme 2. Examples of Pairs of Reactions with the Same or Similar Reactants and Yielding Different Products



of 0.1. The weights were initialized with random numbers that are calculated using the mean and the standard deviation of each variable in the input data set as parameters. For the selection of the winning neuron, the minimum Euclidean distance between the input vector and neuron weights was used. The training was performed over 50−100 cycles, with the learning span and the learning rate linearly decreasing until 0. The Kohonen SOM was implemented with in-house-developed software based on JATOON Java applets.[45] To overcome fluctuations induced by the random factors influencing the training, five independent SOMs were trained with the same objects, generating an ensemble of SOMs, and the final classifications were obtained by majority voting of the individual maps.

Random Forests.[23,46] A random forest is an ensemble of unpruned classification trees created by using bootstrap samples of the training data and random subsets of variables to define the best split at each node. It is a high-dimensional nonparametric method that works well on large numbers of variables. The predictions are made by majority voting of the individual trees. It has been shown that this method is very accurate in a variety of

(45) Aires-de-Sousa, J. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 167−173.
(46) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

Table 1. Classification of Photochemical Reactions by Kohonen SOMs and Random Forests from NMR Data Using Different Fuzzification Parameters

machine learning method	fuzzification width (ppm)	% correct predictions[a] training set	% correct predictions[a] test set
SOM 13 × 13	0.2	99 (97, 96, 98, 97, 96)	88 (76, 79, 81, 76, 81)
	0.3	99 (96, 91, 93, 97, 97)	86 (76, 83, 81, 76, 86)
	0.4	99 (95, 96, 97, 95, 97)	86 (76, 83, 81, 86, 83)
SOM 14 × 14	0.2	99 (98, 95, 96, 97, 95)	93 (79, 79, 81, 83, 88)
	0.3	99 (96, 92, 94, 98, 96)	86 (76, 81, 86, 76, 76)
	0.4	99 (97, 98, 97, 97, 95)	86 (86, 81, 81, 81, 81)
SOM 15 × 15	0.2	99 (96, 97, 98, 99, 95)	93 (81, 81, 81, 81, 86)
	0.3	99 (95, 99, 99, 97, 98)	88 (79, 81, 79, 81, 74)
	0.4	99 (99, 95, 97, 97, 97)	93 (83, 81, 86, 86, 86)
RF	0.2	89[b]	81
	0.3	85[b]	81
	0.4	85[b]	83
RF (classes A−D, F, G repeated)	0.2		86
	0.3		83
	0.4		83

[a] In the SOMs rows, the results were obtained by consensus prediction from an ensemble of five SOMs (within the parentheses are the results for the five individual SOMs). [b] Predictions obtained by out-of-bag estimation (internal cross-validation).

applications.[46] Additionally, the performance is internally assessed with the prediction error for the objects left out in the bootstrap procedure. In this work, RFs were grown with the R program version 2.0.1,[47] using the randomForest library,[48] and were used to classify the reactions from the NMR reaction descriptor, the difference between the ^{1}H NMR spectra of the product and reactants. The number of trees in the forest was set to 1000, and the number of variables tested for each split was set to default (square root of the number of variables).

RESULTS AND DISCUSSION

Validation of the Chemical Shifts Predicted by SPINUS. Although SPINUS has been tested with large, diverse data sets,[21,49] we validated the predicted chemical shifts for some specific types of structures involved in this study by comparing predictions with experimental chemical shifts for some reactants and products in our data set of photochemical reactions. The experimental chemical shifts were obtained from the SDBS database[43] or from the references associated with the reactions in the SPRESI database.

Comparisons were performed for 349 chemical shifts from 36 molecular structures covering reactants and products of all classes. Assignment of the experimental chemical shifts to individual protons was done on the basis of similarities between the predictions and experimental values. This is acceptable for this study, since the global accordance between simulated and experimental spectra is here more relevant than the accuracy of predictions for individual protons. A mean absolute error (MAE) of 0.24 ppm was obtained for the 349 chemical shifts, which is similar to the results of previous tests.[21,49] Table S1 (Supporting Information) details the results by class of reaction. The predictions were particularly accurate for compounds of reaction classes A, D, and G (MAE of 0.16, 0.18, and 0.21 respectively). A MAE lower than 0.2 ppm was achieved for 60% of the cases.

Table 2. Classification of 26 Reactions on the Basis of Experimental and Simulated NMR Data Combined or on the Basis of Simulated Data Alone by the Ensemble of Five Kohonen SOMs and by the Random Forest Trained with Simulated Data

machine learning method	fuzzification width (ppm)	% correct predictions from experimental + simulated data	% correct predictions from simulated data
SOM 13 × 13	0.2	81	100
	0.3	77	96
	0.4	81	92
SOM 14 × 14	0.2	77	100
	0.3	73	96
	0.4	85	96
SOM 15 × 15	0.2	77	100
	0.3	77	96
	0.4	77	96
RF	0.2	81	85[a]
	0.3	81	77[a]
	0.4	92	85[a]

[a] For the reactions included in the training set, the predictions are from internal cross-validation obtained by out-of-bag estimation.

Classification of Photochemical Reactions from ^{1}H NMR Data. SOMs of different sizes were trained with the reactions of the training set using different fuzzification parameters. The (toroidal) surface of a SOM is illustrated in Figure 1, where a trend of some classes to form well-defined clusters is clear. It is to point out that the learning method does not use the information about the classes of the reactions during the training (unsupervised learning). The reactions of the test set were mapped onto the same surface to illustrate the ability of the map to classify unseen reactions. Class A as well as classes C and D clusters into two regions, and inspection of the reactions activating each region showed that they mainly correspond to different types of substrates. Class B is scattered through several regions of the map,

(47) R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-90005.1-07.-0; http://www.R-project.org (accessed Jan 2006).
(48) Fortran original by Leo Breiman, Adele Cutler, R port by Andy Liaw and Matthew Wiener, 2004; http://www.stat.berkeley.edu/users/breiman/ (accessed Nov 2006).
(49) Da Costa, F. B.; Binev, Y.; Gasteiger, J.; Aires-de-Sousa, J. *Tetrahedron Lett.* **2004**, *45*, 6931−6935.

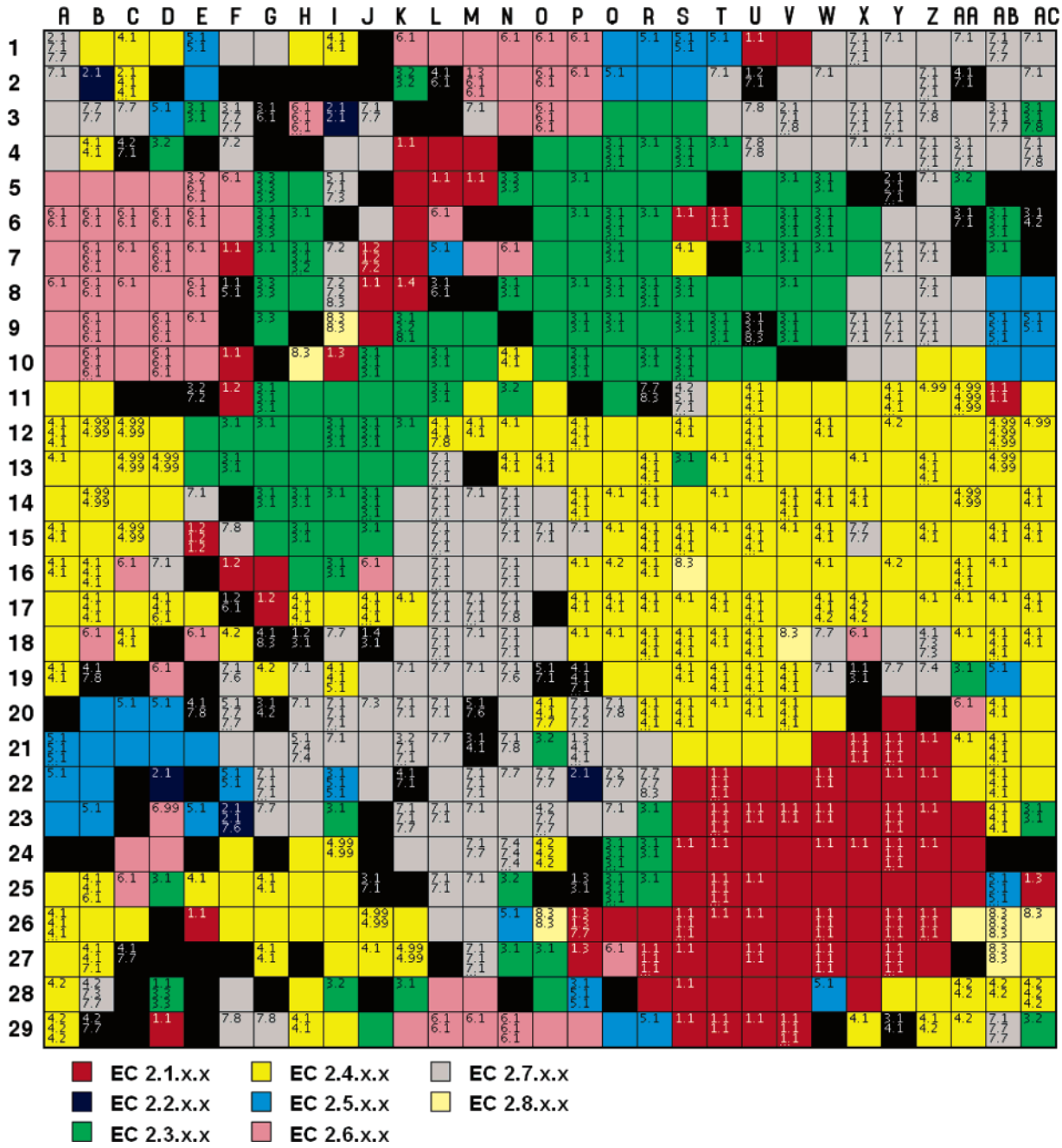


Figure 2. Toroidal surface of a 29 × 29 Kohonen self-organizing map trained with metabolic reactions belonging to EC subclasses 2.1−2.8 on the basis of the ^{1}H NMR descriptor. After the training, each neuron was colored according to the reactions in the training set that were mapped onto it or onto its neighbors. The second and third digits of the EC numbers corresponding to the reactions of the training set are displayed on the neurons they activated. A fuzzification parameter of 0.2 ppm was used for the simulation of NMR spectra.

generally in the vicinity of class E. Class G clusters relatively well with the training set, although the predictions for the test set were correct for only one-half of the cases, even though two of these wrongly classified reactions activated neurons in the neighborhood of a G neuron. The only wrongly classified reaction of class E activated a neuron assigned to class B that was empty with the training set and that is a neighbor of an E neuron. The wrongly classified reaction of class B (B1 in Scheme 2) was classified as D. Inspection of the reactions in the training set that hit the same neuron revealed reactions D2 and D3, which have reactants that are very similar to B1 (Scheme 2). In the test set, another reaction with the same reactants as B1 but giving a different product (D1) also activated this neuron. This shows how the method is sensitive not only to the reaction center but also to the structural environment and, thus, to the structure of the reactants and products.

The percentage of correct classifications obtained for the training and test sets by SOMs are presented in Table 1. Correct predictions could be achieved for 94−99% of the training set and for 81−88% of the test set. The size of the network and the fuzzification parameter have not exhibited a significant influence on the prediction ability.

Consensus predictions involving an ensemble of five independent SOMs considerably improved the predictions both for the training and the test sets (see Table 1). For the training set, only

Table 3. Classification of Metabolic Reactions by Kohonen SOMs and Random Forests

machine learning method	no. reactions, training set	% correct predictions[a]	
		training set	test set (262 reactions)
SOM 25 × 25	911	94 (86, 86, 86, 86, 86)	
	649	96 (89, 89, 91, 89, 88)	73 (67, 66, 66, 66, 66)
SOM 29 × 29	911	96 (89, 89, 89, 90, 89)	
	649	97 (92, 93, 91, 92, 89)	75 (68, 65, 66, 68, 68)
RF	911	84[b]	
	649	84[b]	79

[a] In the SOMs rows, the results were obtained by consensus prediction from an ensemble of five SOMs (within the parentheses are the results for the five individual SOMs). [b] Internal cross-validation obtained by out-of-bag estimation.

one reaction was always wrongly classified: this is reaction D4 of Scheme 2 that has the same reactants as reaction E1 and was mapped into the same neuron. Into this neuron was also mapped reaction E2 (Scheme 2). Predictions for the test set reached up to 93% accuracy. With the ensemble of SOMs of dimension 14 × 14 and fuzzification parameter 0.2 ppm, only one reaction of the test set was wrongly classified (reaction B1 of Scheme 2), and two reactions were undecided (one of class A and one of class G). This ensemble could, thus, overcome most of the problems associated with class G in individual SOMs.

After experimenting with unsupervised learning, we investigated the performance of random forests, which learn in a supervised way. The obtained results are also shown in Table 1. For the training set, the displayed results were obtained in internal cross-validation tests, which rely on the bootstrap procedure employed by the random forests: out-of-bag estimation. Consistency was observed between these results and those obtained for the test set. Most of the wrong predictions were false E. Because class E is the most populated class in the training set, with 38% of the reactions, it was decided to balance the training set by including the reactions of the classes B and G twice, reactions of classes A and D three times, reactions of class F six times and reactions of class C seven times. A slight improvement of the prediction ability for the test set could be achieved, particularly for the experiment with fuzzification parameter 0.2 ppm. In this case, six reactions of the test set were wrongly classified: one of class B, one of class E, and four of class G. Most of the misclassifications were again reactions wrongly classified as class E. Significantly, the problematic reactions of class G (consistently misclassified by SOMs and RFs) have no hydrogen atoms bonded to the atoms of the reaction center.

The voting system of a RF enables the association of a probability to each prediction, corresponding to the proportion of votes obtained by the winning class and by the other classes. For the test set, 29 out of 42 reactions (69%) were predicted with a probability higher than 0.5, and all were correctly predicted.

Validation of the Classification Models with Experimental Chemical Shifts. To assess if the models trained with simulated data could be applied to experimental data, predictions were obtained for a dataset of 26 photochemical reactions in which the simulated chemical shifts for some reactants or products were replaced by experimental values (see Methodology). Both the ensemble of SOMs and the RF were applied (see Table 2). The RF exhibits a higher robustness than the ensemble of SOMs. With the SOMs, the accuracy of the classifications consistently degrades with the inclusion of experimental data in comparison with the experiments based exclusively on simulated data for the same subset of reactions. On the contrary, the accuracy of the RF predictions is quite stable; it is globally the same for simulated data and for the combination of simulated and experimental data.

Classification of Metabolic Reactions from ^{1}H NMR Data. The concept of ^{1}H NMR-based classification of reactions was then explored with a larger, more complex dataset consisting of metabolic reactions catalyzed by transferases. SOMs were trained with a dataset of metabolic reactions on the basis of their ^{1}H NMR descriptor, and neurons were assigned to classes at the end of the training. Figure 2 shows the resulting surface for such a SOM, each neuron colored according to the Enzyme Commission subclass of the reactions activating it, that is, the second digit of the EC number. On the same surface are displayed the second and third digits of the EC numbers corresponding to the reactions hitting each neuron. The map reveals a remarkable clustering of the reactions according to EC numbers, exclusively from the NMR chemical shift-derived descriptors. In a few cases, subregions of the same color (same first three digits of the EC number) could be interpreted in terms of structural features of the reactants or products. For example, reactions with EC numbers 2.6.1 (transaminases, in pink) clustered into two main regions, one around neuron C7, and the other around neuron N1. We observed that the first cluster corresponds to reactions involving transformation of oxoglutaric acid into glutamic acid, whereas the second is associated with transformations of pyruvic acid into alanine. Another example is the region spanned by subclass 2.4 (glycosyltransferases, in dark yellow), where separation is observed for EC numbers 2.4.1 (hexosyltransferases), 2.4.2 (pentosyltransferases, around neuron AA29), and 2.4.99 (transferring other glycosyl groups, around neurons C13 and AB12). In addition, subclass 2.5 (light blue) is clustered into three regions. Reactions activating neurons around position S1 transfer alkyl groups from (*S*)-adenosyl-L-methionine, the cluster around neuron A21 corresponds to reactions transferring alkyl groups from phosphorylated compounds, and the region centered on neuron AB9 was typically activated by *O*-acetyl- or *O*-succinyl-L-homoserine lyases.

In some cases, odd mapping of reactions can be interpreted from their molecular structures. For example, two reactions of subclass 2.7 were mapped on neuron X15, in the middle of the region assigned to subclass 2.4. These two reactions (EC 2.7.7.8) involve the cleavage of a phosphoester bond in RNA and release of a nucleotide. Significantly, cleavage of a sugar−phosphate bond is not typical of subclass 2.7, but is common in the reactions of subclass 2.4 mapped on this region.

Inspection of the map also reveals limitations of the method. Neuron P21 was activated by reactions producing no changes in the ^{1}H NMR spectra. Reactions from different subclasses were mapped on this neuron and its neighborhood, illustrating an expected limitation: when there are no hydrogen atoms in the neighborhood of the reaction center or when the spectra of the products are no different from those of the reactants, the reactions cannot be distinguished or even *described*. Furthermore, in this study, only chemical shifts of hydrogen atoms bonded to carbon atoms were considered. Examples of reactions hitting neuron P21 or its neighbors are

alanine + 2-oxo acid → pyruvate + amino acid (EC 2.6.1.12)

ATP + formate → ADP + formyl phosphate (EC 2.7.2.6)

carbamoyl phosphate + oxamate → orthophosphate + oxalureate (EC 2.1.3.5)

In a series of reactions of the same type, particular features of the spectra can sometimes emphasize differences between substrates over common aspects. This is illustrated by the subclass 2.8. In this study, all reactions of subclass 2.8 belong to subsubclass 2.8.3 (CoA-transferases) and can be formally written as

$$\mathrm{CoA\text{-}S{-}C(=O){-}R^1 + R^2CO_2H \rightarrow CoA{-}S{-}C(=O){-}R^2 + R^1CO_2H}$$

Only slight changes in the ^{1}H NMR spectra result from these reactions, and we observed that the structure of R^1 determined the mapping; reactions with $R^1{=}CH_3$ clustered around neuron AB26 with the others scattered through different regions.

Several Kohonen SOMs, as well as random forests, were trained to classify reactions according to the second digit of the EC number, and their performances were quantitatively evaluated. Table 3 shows the percentage of correct classifications obtained for training and test sets by Kohonen SOMs of sizes 25 × 25 and 29 × 29, and by RFs. A fuzzification parameter of 0.2 ppm was used. In some experiments, the models were trained with the whole dataset; in others, the dataset was first partitioned into a training set and a test set.

With individual SOMs of size 25 × 25 and 29 × 29, correct predictions were achieved for 86% and 89−90%, respectively, of the entire dataset of metabolic reactions. Ensembles of five networks improved the results up to 94 and 96% of correct predictions for SOMs of size 25 × 25 and 29 × 29, respectively. The test set was predicted with 66−67% of accuracy by individual SOMs, and ensembles of five SOMs achieved up to 73% of correct predictions. SOMs of size 29 × 29 yielded slightly better results than SOMs of size 25 × 25. Tables S2 and S3 (Supporting Information) show the confusion matrices obtained for the test set using the best individual SOM and the ensemble of five SOMs of size 29 × 29. The classification performance was worse for subclasses 2.2 and 2.5. These subclasses are subrepresented in the dataset, and most of the misclassifications were as false subclass 2.7.

Employing a random forest as the machine learning method, the predictions for the test set achieved 79% of correct classifications (see Table 3). This result is consistent with the out-of-bag estimation for the entire dataset and for the training set (84% of correct classifications). Again, the subclasses with fewer reactions (2.2 and 2.5) were more difficult to classify, and most of the misclassifications were reactions wrongly classified as subclass 2.7. The confusion matrix for the random forest prediction is available as Table S4 (Supporting Information). The probability associated with each prediction by the RF was again meaningful. In the test set, 165 out of 262 reactions (62%) were predicted with a probability higher than 0.5, and only 7 of these were wrongly classified (4.2%).

The consistent large number of false subclass 2.7 classifications can be explained in part by the very small differences in the ^{1}H NMR spectra of reactants/products predicted by SPINUS for several reactions of subclass 2.7. This was already observed in the SOM of Figure 2, where a number of reactions from different subclasses hit neuron P21 due to their exhibiting null or almost null differences between the spectra of the reactants and products, even though neuron P21 was assigned to subclass 2.7, because most of the reactions by which it was activated belong to that subclass.

CONCLUSIONS

Automatic classification of chemical reactions from differences between ^{1}H NMR spectra of reactants and products was demonstrated with a high level of accuracy for a dataset of photochemical reactions and for a dataset of transferase enzymatic reactions. Ensembles of Kohonen SOMs with consensus predictions allowed for improvement of the results in comparison to individual SOMs. Random forests exceeded SOMs for the metabolic reactions, but not for the photochemical reactions. RFs also allowed associating a meaningful probability to each classification. With the dataset of photochemical reactions, the SOM and RF models were tested with a subset of reactions for which experimental and simulated chemical shifts were combined. Classifications of the same overall quality as those obtained with simulated values alone were obtained by the RF, but SOMs were less robust against the inclusion of experimental data.

The approach is limited by the availability of hydrogen atoms in the neighborhood of the reaction center and by the sensitivity of their chemical shifts to the changes resulting from the reaction.

The results support our proposal of linking reaction and NMR data for automatic reaction classification. They also show the usefulness of SPINUS predictions of NMR data in that context for the generation of training sets and for the estimation of missing experimental data.

Inference of reaction type from NMR experiments can have an application in the assessment of enzymatic function and in the development of biotechnological processes. Classification of the changes in the NMR spectrum of a mixture containing an enzyme and a pool of potential substrates can assist in the exploration of the catalytic ability of the enzyme as well as its versatility.

In most practical situations, a reaction is accompanied by side reactions, which would require that the NMR interpretation system is able to identify a reaction even in the presence of a mixture of reactions. Preliminary results using mixtures of reactions to train neural networks indicate that it is possible to

infer the reaction type when more than one reaction occur simultaneously. Further studies are under way and will be published soon.

ACKNOWLEDGMENT

Diogo A. R. S. Latino acknowledges Fundação para a Ciência e Tecnologia (Ministério da Ciência e do Ensino Superior, Lisbon, Portugal) for financial support under a Ph.D. grant (SFRH/BD/18347). The authors thank InfoChem GmbH (Munich, Germany) for sharing the dataset of photochemical reactions from the SPRESI database. The assistance of Dr. Yuri Binev with the simulation of NMR data is gratefully acknowledged.

SUPPORTING INFORMATION AVAILABLE

Analysis of the estimated ^{1}H NMR chemical shifts in the validation set for photochemical reactions. Confusion matrices for the classification of metabolic reactions in the test set. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review May 28, 2006. Revised . Accepted October 31, 2006.

AC060979S