Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers

Qing-You Zhang and João Aires-de-Sousa*

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

Received July 1, 2005

The automatic classification of chemical reactions is of high importance for the analysis of reaction databases, reaction retrieval, reaction prediction, or synthesis planning. In this work, the classification of photochemical reactions was investigated with no explicit assignment of the reacting centers. Classifications were explored with Random Forests or Kohonen neural networks in three different situations, using different levels of information: (a) pairs of reactants were classified according to the type of reaction they produce, (b) products were classified according to the type of reaction from which they can be synthesized, and (c) reactions were classified from the difference between the descriptors of the product and the descriptors of the reactants. In all cases molecular maps of atom-level properties (MOLMAPs) were used as descriptors. They are generated by a self-organizing map and encode physicochemical properties of the bonds available in a molecule. Correct classification could be achieved for approximately 90% of the 78 reactions in an independent test set.

INTRODUCTION

Chemical reactivity data are available with varied levels of information content. Reactions are entered into databases with molecular structures of reactants and products, and reaction centers may be assigned or not. In other cases, information is available only for the reactants and for some properties of the reaction such as "occurs/does not occur" or rate (important examples are databases of chemical stability under specified storage conditions). Bonds and atoms involved in the reaction (reaction center) may be known or not, as well as the type of the occurring reaction. Studies of chemical reactivity frequently use series of reactants with a common structural motif. Data on chemical reactivity are also implicitly included in some quantitative structure-activity relationship (QSAR) data sets, either because the biological activity consists of a chemical reaction between a compound and a biological target or because the compound undergoes metabolic activation (i.e., it suffers chemical transformations) before producing the measured biological activity.

Models related to chemical reactivity can be built with different goals using different types of data, for example, qualitative or quantitative reaction prediction from the molecular structures of reactants, the suggestion of synthetic methods from the structures of the products, and the classification of reactions from the molecular structures of reactants and products, from the molecular structures of products and the bonds involved in the reaction, or from the molecular structures of the reactants and the bonds involved in the reaction.

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

Naturally, the desired goal and the availability of data put requirements on the representation of structural and reaction data and on the chosen algorithms for learning.

After a period of near inactivity, the topic of automatic classification of chemical reactions[1] is re-emerging, particularly because of the current interest in metabolic reactions.[2] Well-documented methods for reaction classification have been based on (a) physicochemical properties of bonds or atoms at the reaction center,[3−6] (b) variation of physicochemical properties of a specific atom (e.g., oxygen atom) attached to the reaction center,[7] (c) codes of the neighborhoods of the reaction center,[2,8−11] and (d) fragment keys for the reaction center and the substrates.[12] These approaches require atom mapping and the identification of the reaction center. Some also require ranking of the bonds involved in the reaction and a scheme to compare reactions with an unequal number of bonds involved.[6]

Besides the classical approaches to computer-aided reaction prediction and synthesis design,[13] in which the structures of the products are explicitly predicted from the structures of the reactants, or vice versa, other data-driven approaches have appeared related to the processing of chemical reactions data where the focus is on *classification*. Chen and Gasteiger have explored Kohonen neural networks to define a space of chemical reactions on a map, using physicochemical properties of the reaction center as the numerical representation of the reaction.[5,6,14] Their work was followed by that of Satoh et al.[7] The Kohonen map allows for reaction classification, reaction prediction, and the establishment of a reaction similarity metric. Hajduk and co-workers[15] developed a filtering tool based on substructural fragments to classify compounds as reactive or nonreactive toward the human La antigen, a reactivity probe in the high-throughput screening of protein targets. Satoh et al.[16] proposed the FRAU

10.1021/ci0502707 CCC: $30.25
Published on Web 10/20/2005

method to numerically characterize a field around an atom in a molecule, which is based on electrostatic and steric interactions with a pseudoreactant. This method was applied to the classification of reagents by self-organizing maps (SOMs) in terms of their roles in reactions.[17]

Chemical reactivity, being related to the ability to make and break bonds, is primarily determined by properties of the bonds available in a molecule. Gasteiger et al.[18] proposed that seven empirical physicochemical properties are particularly relevant for representing bonds and for modeling chemical reactivity: σ electronegativity, π atomic charge, total atomic charge, bond polarity, mean bond polarizability, resonance stabilization, and bond dissociation energy. To use all that information for an entire molecule, and at the same time having a fixed-length representation, we propose to map all the bonds of a molecule into a fixed-length 2D self-organizing map—a *MOLMAP* (molecular map of atom-level properties).

A SOM must be trained beforehand with a diversity of bonds from different structures (each bond described by the seven bond properties). Then, all the bonds of one molecule are submitted to the trained SOM, each bond activates one neuron, and the pattern of activated neurons is a map of the reactivity features of that molecule (MOLMAP)—a fingerprint of the bonds available in that structure.

The MOLMAP descriptors can be directly used for data mining or QSAR studies related to chemical reactivity, in situations involving different types of reaction sites in a single data set, more than one reaction site in a single structure, or unknown reaction sites. The first part of this paper illustrates such an approach with the application of MOLMAP descriptors of reactants, or of products, to reaction prediction using a data set of 356 photochemical reactions classified into seven types. Random Forests (RFs) were trained on the basis of the MOLMAPs of the reactants, or of the products, to predict the type of the occurring reaction.

A method somewhat related to the MOLMAP method is the comparative molecular surface analysis[19] that maps points of the molecular surface into a self-organizing map, assigns numbers to the neurons according to the electrostatic potential of the points, and compares the maps of different molecules.[20] Points of the molecular surfaces are submitted to the SOM in the form of 3D Cartesian coordinates, which requires alignment of the 3D models, or of the maps, before comparison.

In the course of our studies with MOLMAPs, it was recognized that the difference between the MOLMAPs of the products of a reaction and the MOLMAPs of the reactants of the same reaction could be interpreted as a MOLMAP *of the reaction*. The MOLMAP of a molecule is the pattern of neurons that are activated by the bonds existing in that molecule. Bonds far apart from the reacting center are mostly unchanged during the reaction, exhibiting the same physicochemical properties and, thus, activating the same position (neuron) of the MOLMAP both in the reactants and in the products. Therefore, the difference map (MOLMAP of the reaction) gets a zero value at that neuron. The pattern of neurons in the MOLMAP of the reaction with nonzero values relates to the bonds of the reactants that break or change properties and to the bonds of the products that are made or changed in the reaction. The former lead to negative values, while the latter lead to positive values. In this way, a representation of the reaction is achieved *without previous assignment of the reaction center*. The idea of subtracting a matrix of the products from a matrix of the reactants to define a matrix of the reaction was proposed by Dugundji and Ugi in the mid-1970s.[21] However, Ugi reaction matrices are fundamentally different in several aspects, namely, in their connectivity matrix essence, their requirement of atom-to-atom mapping, and their variable size depending on the number of atoms in the molecules.

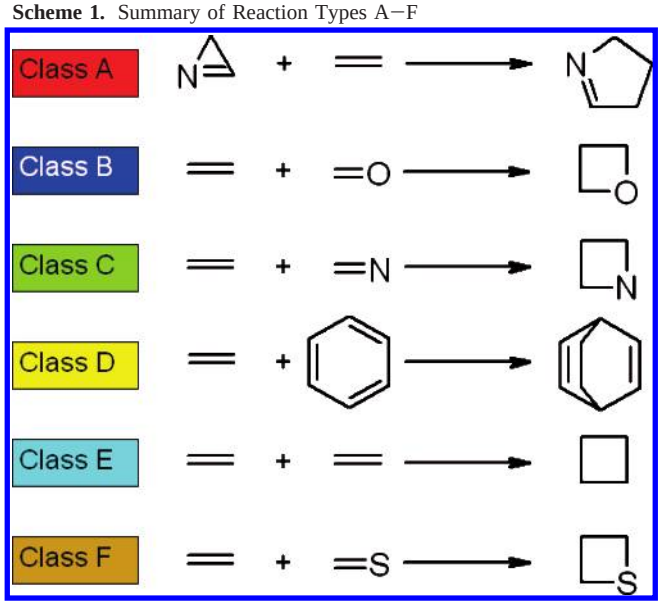
In the second part of this paper, the photochemical reactions in our data set were encoded with the MOLMAPs of the reactions. The ability of this code to classify reactions was assessed by an unsupervised machine-learning method (Kohonen self-organizing map) as well as a supervised method (Random Forests).

METHODOLOGY

The experiments here described required two major steps, the generation of the descriptors (molecular and reaction descriptors) and the development of predictive models. Generation of the descriptors, the so-called MOLMAPs, was based on a Kohonen SOM.[22] For training this map, each object of the training set is a chemical bond, represented by seven empirical physicochemical properties. This training set comprises bonds from a diversity of structures. Once trained, the map is used to obtain molecular descriptors; all the bonds of one molecule are submitted to the map, and the resulting pattern of activated neurons (the MOLMAP) is the descriptor of the molecule. The second step consists of establishing relationships between MOLMAPs and reactivity properties (class of reaction). Models were developed to (a) predict which type of reaction takes place from the MOLMAPs of the reactants, (b) predict which type of reaction should take place from the MOLMAPs of the product, and (c) classify a reaction from the MOLMAPs of its reactants and product. The second step is independent of the first, even if, in one case, Kohonen self-organizing maps were also used for the second step. It must be emphasized that, in that case, while in the first step each object is a bond, in the second step each object is a reaction.

Data Set. A set of 356 reactions, tagged as involving "irradiation" and consisting of two reactants and one product, was extracted from the SPRESI database (InfoChem GmbH, Munich, Germany). Reactions differing only in stereochemical features were considered as duplicates and were included only once. The reactions were manually classified into seven types: [3+2] photocycloaddition of azirines to C=C (18 reactions, class A), [2+2] photocycloaddition of C=C to C=O (37 reactions, class B), [2+2] photocycloaddition of C=N to C=C (13 reactions, class C), [4+2] and [4+4] photocycloaddition of olefins to carbon-only aromatic rings (20 reactions, class D), [2+2] photocycloaddition of C=C to C=C (78 reactions, class E), [2+2] photocycloaddition of C=C to C=S (27 reactions, class F), and a seventh group with the remaining reactions (163 reactions, class G). Scheme 1 shows the first six types of reactions. The data set was divided into a training set consisting of 278 reactions (14 of type A, 28 of type B, 10 of type C, 16 of type D, 62 of type E, 22 of type F, and 126 of type G) and a test set with 78 reactions (4 of type A, 9 of type B, 3 of type C, 4 of type D, 16 of type E, 5 of type F,

Scheme 1. Summary of Reaction Types A−F



and 37 of type G). This partition was done randomly, but it was assured that both sets cover the whole diversity space of the reactions data set.

Training of a Kohonen Self-Organizing Map with Bonds. Kohonen SOMs can be used for the reduction of multidimensional objects to 2D. In this study, we used a SOM to reduce to 2D the dimension of chemical bonds, represented by seven bond properties calculated by PETRA 3.20:[23] resonance stabilization, the difference between the σ electronegativity of the two atoms, the difference between the total charge of the two atoms, the difference between the π charge of the two atoms, the mean bond polarizability, bond dissociation energy, and the sum of σ charges shifted over all iterations.[24,25] As some properties depend on the orientation of the bond, each bond was represented twice (as A−B and B−A). For all the properties to be equally important, each of them was *z*-normalized on the basis of the whole data set.

SOMs learn by unsupervised training, revealing similarities between objects (bonds). A Kohonen SOM consists of a grid of so-called neurons, each containing as many elements (weights) as the number of input variables. Here, the input variables are the seven properties of bonds (Figure 1). Before the training starts, the weights take random values. During the training, each individual bond is mapped into the neuron that contains the most similar weights compared to its properties. This is the central neuron, or winning neuron. It is said that the winning neuron is *excited* by the bond, and its weights are then adjusted to make them even more similar to the properties of the presented bond. Not only does the winning neuron have its weights adjusted but also the neurons in its neighborhood. The extent of adjustment depends, however, on the topological distance to the winning neuron: the closer a neuron is to the winning neuron, the larger is the adjustment of its weights. The objects of the training set are iteratively fed to the map, the weights corrected, and the training is stopped when a predefined number of cycles is attained. A trained Kohonen SOM reveals similarities in the objects of a data set in the sense that similar objects (similar bonds) are mapped into the same or closely adjacent neurons.

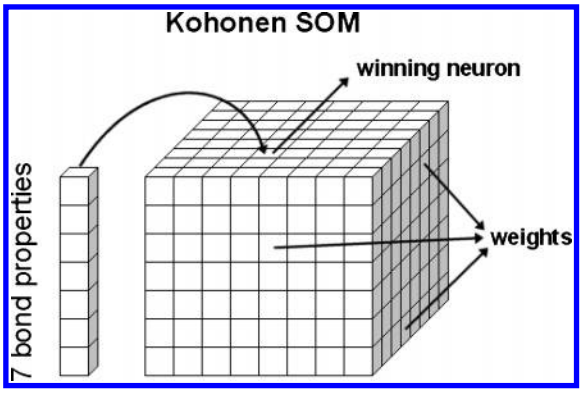


Figure 1. Representation of a Kohonen SOM for processing chemical bonds. Every small box of the block represents a weight. The Kohonen SOM is trained by the iterative presentation of objects (bonds described by seven properties).

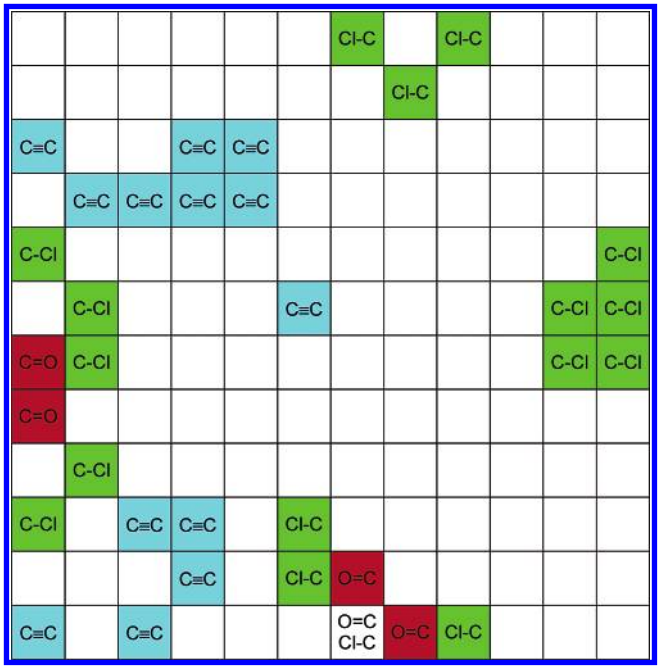


Figure 2. Representation of the neurons in the SOM activated by C═O, C≡C, and C−Cl bonds.

SOMs of sizes 12 × 12 or 15 × 15 were trained with a random subset of 4999 bonds extracted from the reactions data set. SOMs were implemented with in-house-developed software based on JATOON Java applets.[26,27] Figure 2 illustrates the activation of neurons in a trained map by carbon−halogen bonds, C≡C bonds, and C═O bonds, showing how different regions of the surface are activated by different types of bonds.

Molecular Descriptors (Molecular MOLMAPs). A representation of the set of bonds existing in a molecule can be obtained by mapping the bonds of that molecule on the SOM previously trained with a diversity of bonds. The pattern of activated neurons can be interpreted as a fingerprint of the available bonds in the molecule, and it was used as a molecular descriptor (MOLMAP). For numerical processing, each neuron got a value equal to the number of times it was activated by bonds of the molecule. The map was then transformed into a vector by concatenation of the columns. To account for the relationship between the similarity of bonds and proximity in the map, a value of 0.3 was added to each neuron multiplied by the number of times a neighbor

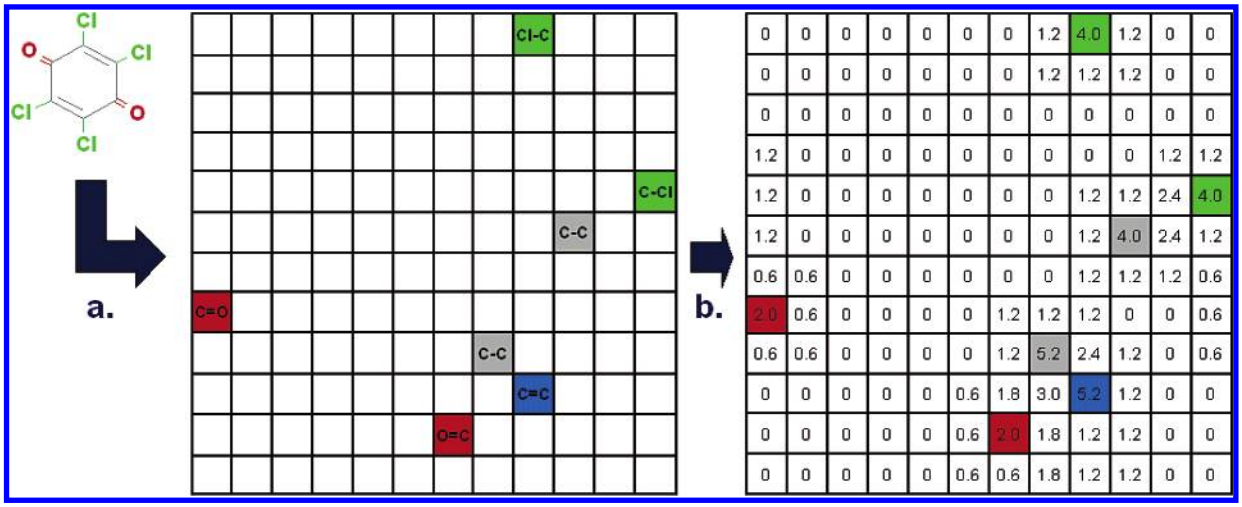


Figure 3. Generation of the MOLMAP descriptor for a simple molecule. (a) Submission of bonds to a previously trained SOM. (b) Numerical processing of the pattern of activated neurons.

was activated by a bond. If an empty neuron is a direct neighbor of more than one activated neuron, its value is the sum of the contributions from each activated neuron. Figure 3 illustrates the generation of the MOLMAP for a molecule.

Reaction Descriptors (Reaction MOLMAPs). A chemical reaction is defined by the changes operated in the chemical bonds of the reactants, leading to the formation of the products. These changes can be represented by the difference between the MOLMAPs of the products and the MOLMAPs of the reactants. Here, we propose to represent a chemical reaction by this difference. If the reaction involves more than one reactant, the MOLMAPs of all the reactants must be overlapped and numerically summed, and if the reaction results in more than one molecule, the MOLMAPs of the products must be processed in the same way. The MOLMAP *of the reaction* is obtained by subtracting the MOLMAP of the reactants from the MOLMAP of products. A step-by-step illustration of the procedure is provided as Supporting Information.

Random Forests.[28,29] 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. A prediction is made by majority vote of the individual trees. It has been shown that the method is extremely accurate in a variety of applications.[29] Additionally, the performance is internally assessed with the prediction error for the objects left out in the bootstrap procedure. The method quantifies the importance of a variable by the increase in misclassification occurring when the values of the variable are randomly permuted. In this study, RFs were grown with the R program, version 2.0.1,[30] using the randomForest library.[31] They were used to predict the class of the reaction from the MOLMAP of the reactants, to suggest a class of reaction leading to a product represented by its MOLMAP, and to classify the reaction from the reaction MOLMAP.

Reaction Prediction from the Structures of the Reactants. MOLMAP descriptors were calculated for all the reactants of the reaction data set. As all the reactions in this study have two reactants, the MOLMAPs of the two reactants were summed. RFs were grown with a training set consisting of the MOLMAPs of the reactants as the independent variables and the manually assigned class of the reaction as the dependent variable. In these experiments, no information was used concerning bonds or atoms involved in the reaction or the structure of the resulting products. Among the 356 reactions, some reactions have the same two reactants and belong to the same class but the products are different. When only the MOLMAPs of the reactants are used, such objects are exactly the same, and only one is kept. After this filtering, 333 reactions were left. The training set was composed of 260 reactions, and the test set consisted of 73 reactions. The number of trees in a RF was set to 1000, and MOLMAPs of dimensions 12×12 and 15×15 were tested.

Reaction Suggestion from the Structure of the Product. MOLMAP descriptors were calculated for all the products of the reaction data set. RFs were grown with a training set consisting of the MOLMAPs of the products as the independent variables and the manually assigned class of the reaction as the dependent variable. In these experiments, no information was used concerning the bonds or atoms involved in the reaction or the structure of the reactants. MOLMAPs of dimensions 12×12 and 15×15 were tested.

Reaction Classification from the Structures of the Reactants and Product. The molecular structures of the products and the reactants were encoded by their MOLMAPs, and each reaction was encoded by the difference between the MOLMAP of the product and the MOLMAP of the reactants. Two machine learning techniques, RFs and SOMs, were used to classify reactions on the basis of their MOLMAPs. While SOMs are trained with unsupervised learning, RFs are trained with supervised learning.

Kohonen SOMs were trained as described above, but now using reactions as the objects, described by their MOLMAPs. After the training, the whole training set was mapped on the surface, and each neuron was assigned to a class of reactions depending on the majority of reactions that activated the neuron. When a majority could not be obtained, the neuron was classified as undecided. To overcome fluctuations induced by the random factors influencing the

Table 1. Random Forest Prediction of Reaction Class from the MOLMAPs of the Reactants

code length	% correct predictions: training set[a]	% correct predictions: test set[b]
144 (12 × 12 MOLMAPs)	80	82
225 (15 × 15 MOLMAPs)	82	82
144 (12 × 12 MOLMAPs; classes A−F repeated)		85 (87)
225 (15 × 15 MOLMAPs; classes A−F repeated)		88 (90)

[a] Internal cross-validation obtained by out-of-bag (OOB) estimation. [b] Within the parentheses are the results excluding the two reactions having the same reactants but belonging to different classes.

training, five independent SOMs were trained, generating an ensemble of SOMs. Predictions were obtained by majority vote of the individual maps.

RFs were grown with a training set consisting of the MOLMAPs of the reactions as the independent variables and the manually assigned class of the reaction as the dependent variable. In these experiments, no information was used concerning the assignment of reacting centers (bonds or atoms involved in the reaction) or atom-to-atom mapping.

RESULTS AND DISCUSSION

Reaction Prediction from the Structures of the Reactants. The results obtained by the RFs with MOLMAPs of varying sizes are shown in Table 1. The results obtained in internal cross-validation tests with the training set are consistent with those obtained for the external test set. The internal cross-validation relies on the bootstrap procedure employed by the RFs. Each tree is grown with only a subset of objects, and predictions are obtained for the objects left out. The global result for objects left out by the 1000 trees is used for assessing the quality of the trained forest—out-of-bag (OOB) error estimation. A higher resolution of the MOLMAP only slightly improved the classifications.

As class G is over-represented, and many erroneous predictions are false G predictions, it was decided to balance the data set by including classes A−F twice. An increased number of correct predictions for the test set could thus be achieved.

These are good results, particularly considering the limitations inherent to this approach, to the data set, and to the exclusive use of empirical calculated physicochemical properties. Two reactants can, in principle, react in several different ways depending on the experimental conditions. Because all the reactions in our data set are promoted by irradiation, we assumed that some homogenization of the experimental conditions was in place and that the same functional groups in the reactants generally yield the same class of reaction. However, this is not always true. The data set includes distinct reactions where the same functional groups react in different ways. This occurs as a result of the inclusion of different specific reaction conditions and also the inclusion of both main reactions and side reactions. We found the extreme situation in two cases where the same reactants give rise to different reactions of different classes (reactions D1, E1, G1, B2, and D2; Scheme 2). Such cases obviously create noise in the model and limit the quality of the predictions. Reactions D1, E1, and D2 were in the training set, while reactions G1 and B2 were in the test set. Removal of these cases from the test set naturally improved the percentage of correct predictions (Table 1). It is noted that the data set in this study specifies that a certain pair of reactants react in a certain way, but there is no information guaranteeing that a certain class of reaction cannot occur for that pair of reactants.

Scheme 2. Two Cases in Which the Same Two Reactants Produce Different Types of Reactions

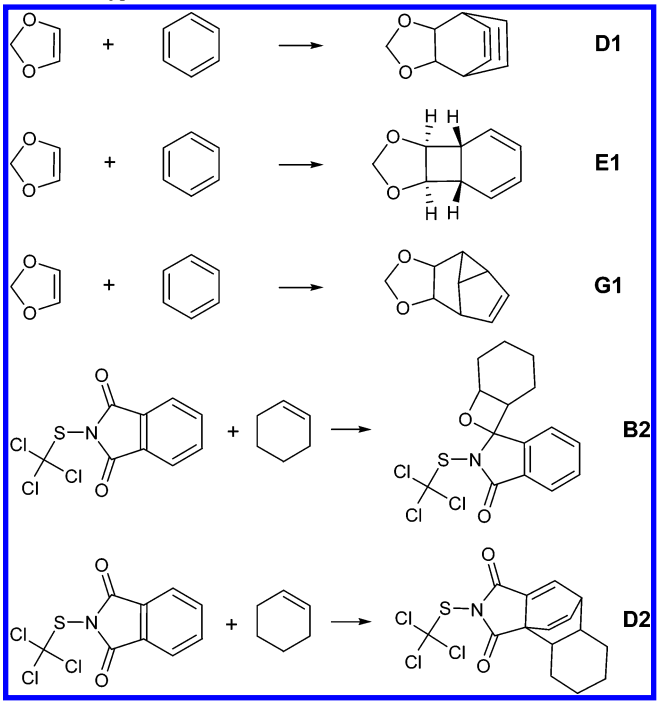


Table 2. Confusion Matrix for the Predictions of Reaction Type from the 15 × 15 MOLMAPs of the Reactants Obtained by a Random Forest for the Test Set

	A	B	C	D	E	F	G	class error
A	4							0%
B		8						0%
C			2					0%
D				3			1	25%
E					13		2	13%
F						4	1	20%
G					1	2	30	9%

The confusion matrix for the test set (without reactions G1 and B2) using a code length of 225 and classes A−F repeated is shown in Table 2. Classes D and F gave the largest errors, and these are significantly the least represented classes.

The voting system of the forest enables the association of a probability to each prediction. This is the proportion of votes obtained by the winning class and by the other classes. We observed that 52 out of 73 reactions of the test set were predicted with a probability higher than 0.5. Only 3 of these 52 reactions were wrongly predicted; one is reaction B2. The other two belong to class G but were predicted as class F. Both of them comprise C═S bonds in the reactants and are cycloaddtitions of C═S to C≡C groups.

Inspection of the importances determined by the RFs for each variable and for each class revealed that no single variable (or group of variables) has a far higher importance than the others. This suggests that the model is based mostly on the cooperative contribution of many variables. However,

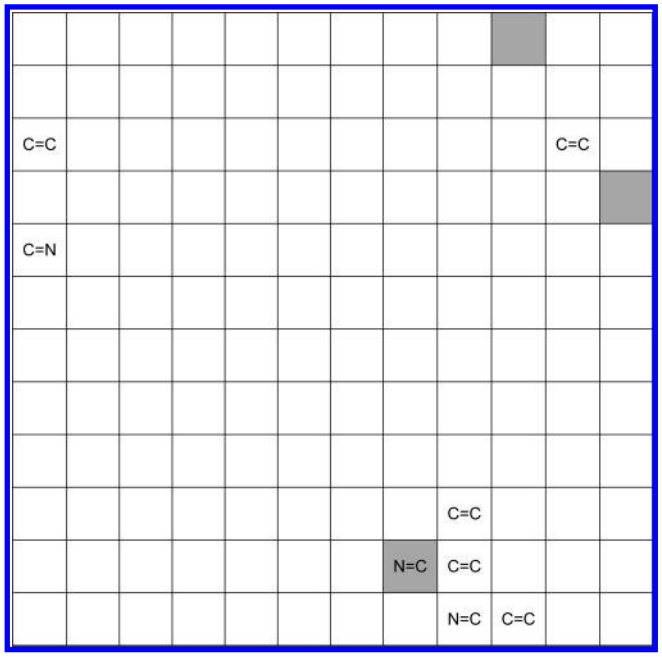


Figure 4. Representation of the three most important components of the MOLMAP for class C (colored in gray) and their neighbors related to C═C and C═N bonds.

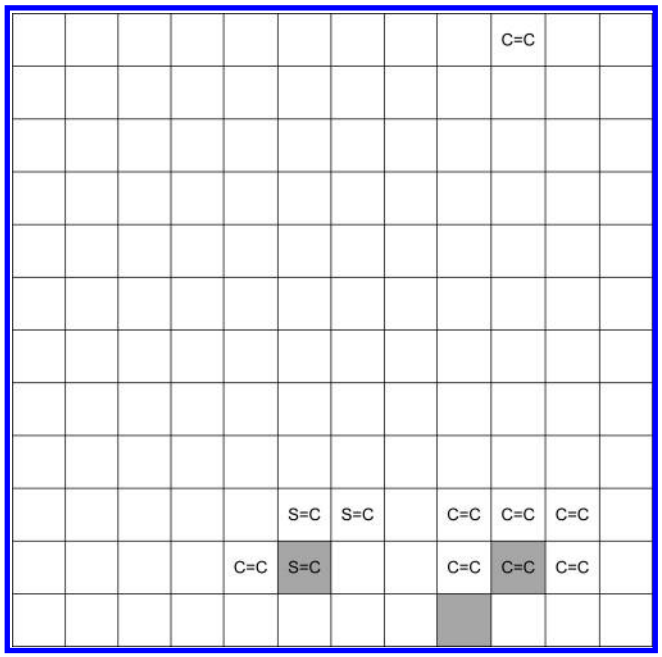


Figure 5. Representation of the three most important components of the MOLMAP for class F (colored in gray) and their neighbors related to C═C and C═S bonds.

some expected observations could be made concerning the highest ranked variables. For example, the three most important variables for class C (photocycloaddition of C═N to C═C) are related to MOLMAP neurons occupied by C═N and C═C bonds, Figure 4. Similar results were observed for class F (photocycloaddition of C═C to C═S), Figure 5.

Suggestion of Reaction Type Leading to the Structure of the Product. Several synthetic strategies are usually possible for a given molecule. However, in the data set of reactions used in this study, the substructure produced by each of the seven classes of reactions is unique and does not appear in the structure of products belonging to other classes for 99% of the cases. It is, therefore, reasonable to classify a product according to the type of reaction from which it can be synthesized. The RF trained with the MOLMAPs of the products as the independent variable, and the assigned class of the reaction as the dependent variable, gave predictions consistent with the manually assigned class for 88% and 91% of the test set, respectively, using MOLMAPs of dimensions 12 × 12 and 15 × 15. Internal cross-validation for the training set gave 80% and 83% of consistent predictions, for the two dimensions of MOLMAPs, respectively. The training set of 278 reactions and the test set of 78 reactions were used. The confusion matrix for the test set using a code length of 225 is shown in Table 3. Again, the least represented classes, D and F, gave the largest errors. Probabilities of the predictions gave a robust indication of the consistency with the assigned classes. For the 58 predictions (out of 78) obtained with a probability above 0.5, consistent classification was obtained for 98% of the cases.

Table 3. Confusion Matrix for the Suggestion of Reaction Type from the 15 × 15 MOLMAPs of the Products Obtained by a Random Forest for the Test Set

	A	B	C	D	E	F	G	class error
A	4							0%
B		8			1			11%
C			3					0%
D				3			1	25%
E				1	13		2	19%
F						4	1	20%
G		1					36	3%

The results obtained for the prediction of a reaction from the structures of the reactants, and for the suggestion of reaction type from the structure of the product, demonstrate the ability of the MOLMAPs−RF system to learn the reactivity behavior of chemical compounds described in a database of experimental examples. It is emphasized that this approach does not explicitly encode substructures.

Classification of Reactions from the Structures of Reactants and Product. In the experiments described up to here, information was used that was about only one side of the reaction equation (either the reactants side or the product side) and the class of the reaction. RFs were trained to establish relationships between the molecular descriptors of the reactants (or product) and the type of reaction to which they are related. Now, we describe experiments that use information about both sides of the reaction equation (differences between the structures of the reactants and the product) and the class of the reaction. In this case, we decided to investigate classification by an unsupervised method, in addition to a supervised method, as every component of the reaction descriptor (reaction MOLMAP) describes a change between the structures of the product and reactants, thus, containing potentially useful information about the reaction.

Reaction MOLMAPs for the 278 reactions of the training set were used to train a 17 × 17 Kohonen SOM. After the training, the neurons were assigned classes according to the reactions that activate them, and these assigned classes are used to make predictions for new reactions when they are submitted to the SOM. During the training, the information

Table 4. Classification of Reactions by Kohonen SOMs from the MOLMAPs of the Reactions

code length	% correct predictions[a]	
	training set	test set
144 (12 × 12 MOLMAPs)	96 (88, 85, 92, 89, 87)	83 (71, 62, 71, 77, 72)
225 (15 × 15 MOLMAPs)	97 (89, 86, 86, 90, 91)	86 (71, 73, 73, 72, 73)

[a] Within the parentheses are the results for the five individual SOMs of the ensemble.

Table 5. Confusion Matrix for the Classification of Reactions from the 15 × 15 Reaction MOLMAPs Obtained by an Ensemble of Five SOMs for the Test Set

	A	B	C	D	E	F	G	undecided	class error
A	4								0%
B		7		1			1		22%
C			3						0%
D				4					0%
E		1		1	12			2	25%
F					1	3	1		40%
G	1				1	1	34		8%

concerning classes is not used; that is, the reactions are mapped into neurons exclusively on the basis of their MOLMAP descriptors—unsupervised learning. To reduce the impact of fluctuations derived from the random values of the weights at the outset of the training, five SOMs were trained independently, and predictions were obtained by majority voting.

Results obtained with MOLMAPs of dimensions 12 × 12 and 15 × 15 are displayed in Table 4. The results show that the reaction MOLMAPs allow for the definition of a reaction space with a topology related to the proposed classes of the reactions. The advantage of using an ensemble of five SOMs is apparent: no single SOM gave better results than the ensemble. The confusion matrix for the test set in the experiment with MOLMAPs of dimension 15 × 15 is shown in Table 5. A reaction is considered undecided if no class obtained more votes than any of the others. After removing class G from the training and test sets, an ensemble of five 13 × 13 Kohonen SOMs correctly classified 90% of the reactions (represented by MOLMAPs of dimension 15 × 15) in the test set, which indicates that class G does not create significant noise in the model.

The surface of a 13 × 13 Kohonen SOM is displayed in Figure 6, with neurons colored according to the classes of the reactions in the training set mapped onto them. It reveals good clustering by the class of reaction. The reactions of the test set were mapped onto the trained SOM and are represented by the labels of their classes. The surface has a toroidal topology, implying that neurons at the top row are neighbors of neurons at the bottom row, and neurons at the left column are neighbors of those at the right column. A trend is observable for classes to occupy a homogeneous region. Class D (cycloadditions of olefins to aromatic rings) presents two clusters corresponding to two types of aromatic compounds; in the small cluster, the aromatic substrates are phthalic acid derivatives. The two classes D and E, with only carbon atoms at the reacting center, were mapped in the vicinity of each other. Class C (cycloaddition of C=C to C=N) also occupies two distinct regions, one corresponding to substrates with the nitrogen atom of the reacting center bonded to another nitrogen atom and the other with the

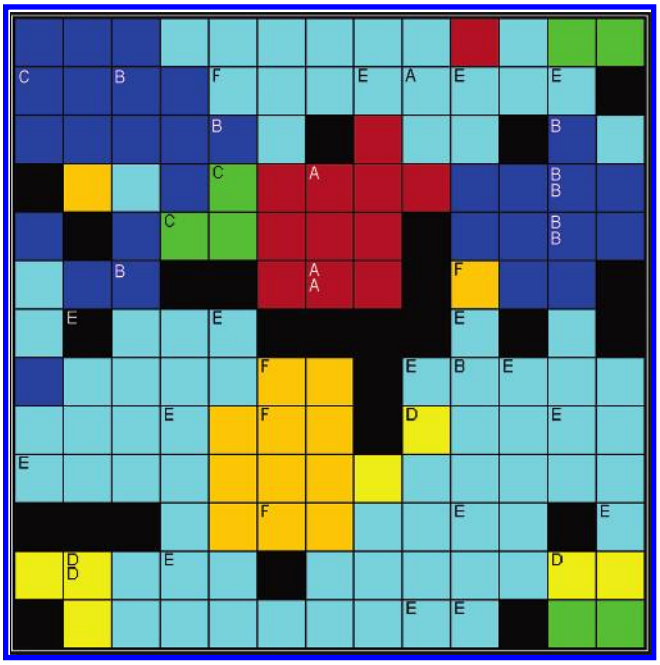


Figure 6. Toroidal surface of a SOM trained with MOLMAPs of reactions belonging to classes A−F. After the training, each neuron was colored according to the reactions in the training set that were mapped onto it. If a neuron was not activated by any reaction, it was classified according to the class of its neighbors by majority voting. Colors were defined in Scheme 1. The reactions of the test set were mapped onto the trained SOM and are represented by the label of their classes. Neurons with undecided assignments were colored in black.

nitrogen atom bonded to an oxygen atom. Class F is the less-clustered class, with one large cluster and two distant individual neurons, one of them separated from the cluster by undecided neurons and the other corresponding to reactions involving a common substrate. Class F is the class with the worst predictions in Table 5. Five reactions of the test set were wrongly classified by this SOM—one reaction from classes A, B, C, E, and F. It can be seen that the wrongly classified reactions of classes A, C, and E activated neurons that are neighbors of the true regions for their classes. The wrongly classified case of class B corresponds to reaction B2 in Scheme 2. It activated a position that is in the neighborhood of the neuron activated by reaction D2 and by other [4+2] cycloadditions of cyclohexene to phthalic acid derivatives. It is also in the neighborhood of two undecided neurons and two neurons away from its true region.

The results indicate that the MOLMAPs of the reactions are influenced by relatively large substructures of the substrates around the reaction center. This was confirmed by the observation that unsupervised learning (ensemble of SOMs) with the MOLMAPs of the *reactants* produced a reasonable percentage of correct predictions for the test set (63%). This is, however, significantly inferior to the performance of the MOLMAPs of the *reactions* (86%), which shows that the MOLMAPs of the reactions focus, indeed, on the structural changes resulting from the reaction.

After the results with unsupervised learning revealed a natural relationship between reaction MOLMAPs and reaction classes, it was investigated if a supervised method could refine the classification accuracy. The results obtained by RFs for the classification of reactions from their MOLMAPs

Table 6. Random Forest Classification of Reactions from the MOLMAPs of the Reactions

code length	% correct predictions	
	training set[a]	test set
144 (12 × 12 MOLMAPs)	84	90
225 (15 × 15 MOLMAPs)	86	91

[a] Internal cross-validation obtained by OOB estimation.

Table 7. Confusion Matrix for the Classification of Reactions from the 15 × 15 Reaction MOLMAPs Obtained by a Random Forest for the Test Set

	A	B	C	D	E	F	G	class error
A	4							0%
B		7			1		1	22%
C			2				1	33%
D				4				0%
E					15		1	6%
F						3	2	40%
G					1		36	3%

are presented in Tables 6 and 7. The accuracy of classifications was enhanced comparing to the unsupervised method. An analysis of the probabilities associated with the predictions showed that 60 out of the 78 reactions of the test set obtained a probability above 0.5, and only two reactions were wrongly predicted among these. Class F was, again, the most problematic class.

CONCLUSION

Three types of problems related to the classification of chemical reactions were approached using MOLMAP descriptors of bond physicochemical features: the prediction of the class of reaction could be achieved from the MOLMAPs of the reactants, meaningful suggestions of a reaction type leading to a given product could be achieved from the MOLMAP of the product, and the difference between the MOLMAPs of the product and reactants could be used to classify the reaction exclusively on the basis of physicochemical properties, and with no assignment of the reaction center. The MOLMAPs of reactions were influenced not only by the bonds of the reaction center but also by other bonds of the substrates, around the reaction center. The methods here presented have potential applications in automatic reaction prediction systems, computer-aided synthesis design, and particularly in the analysis and data mining of databases of chemical reactions. The encoding of chemical reactions by the difference of products' descriptors and reactants' descriptors opens the way to many possibilities far beyond the use of MOLMAP descriptors. It allows for an easy exploration of reaction databases making use of a large diversity of molecular representations.

ACKNOWLEDGMENT

The authors thank InfoChem GmbH (Munich, Germany) for making available a subset of reactions from the SPRESI database and Molecular Networks GmbH (Erlangen, Germany) for access to the PETRA software package. Gonçalo Carrera is acknowledged for exploratory experiments with MOLMAPs. Z.Q.Y. acknowledges Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) for a postdoctoral grant under the POCTI program (SFRH/BPD/14476/2003).

Supporting Information Available: A step-by-step illustration of the procedure to construct the MOLMAP descriptor for a reaction. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.
(2) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational Assignment of the EC Numbers for Genomic-scale Analysis of Enzymatic Reactions. *J. Am. Chem. Soc.* **2004**, *126*, 16487−16498.
(3) Chen, L.; Gasteiger, J.; Rose, J. R. Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon−Hydrogen Bonds to Carbon−Carbon Double Bonds. *J. Org. Chem.* **1995**, *60*, 8002−8014.
(4) Rose, J. R.; Gasteiger, J. HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 74−90.
(5) Chen, L.; Gasteiger, J. Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network. *J. Am. Chem. Soc.* **1997**, *119*, 4033−4042.
(6) Sacher, O. *Classification of Organic Reactions by Neural Networks for the Application in Reaction Prediction and Synthesis Design*. Ph.D. Thesis, University of Erlangen−Nuremberg, Erlangen, Germany, 2001. http://www2.chemie.uni-erlangen.de/services/dissonline/data/dissertation/Oliver_Sacher/html/
(7) Satoh, H.; Sacher, O.; Nakata, T.; Chen, L.; Gasteiger, J.; Funatsu, K. Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210−219.
(8) CLASSIFY. The InfoChem. Reaction Classification Program. V.2.5. http://www.infochem.de/eng/downloads/Classify.pdf
(9) Fujita, S. Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205−212.
(10) Hendrickson, J. B. Comprehensive System for Classification and Nomenclature of Organic Reactions. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 852−860.
(11) Tratch, S. S.; Zefirov, N. S. A Hierarchical Classification Scheme for Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 349−366.
(12) Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiau, J. Similarity Searching in the Organic Reaction Domain. *Tetrahedron Comput. Methodol.* **1988**, *1*, 117−128.
(13) Todd, M. H. Computer-aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247−266.
(14) Chen, L.; Gasteiger, J. Organic Reactions Classified by Neural Networks: Michael Additions, Friedel−Crafts Alkylations by Alkenes, and Related Reactions. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 763−765.
(15) Huth, J. R.; Mendoza, R.; Olejniczak, E. T.; Johnson, R. W.; Cothron, D. A.; Liu, Y.; Lerner, C. G.; Chen, J.; Hajduk, P. J. ALARM NMR: A Rapid and Robust Experimental Method To Detect Reactive False Positives in Biochemical Screens. *J. Am. Chem. Soc.* **2005**, *127*, 217−224.
(16) Satoh, H.; Itono, S.; Funatsu, K.; Takano, K.; Nakata, T. A Novel Method for Characterization of Three-Dimensional Reaction Fields Based on Electrostatic and Steric Interactions toward the Goal of Quantitative Analysis and Understanding of Organic Reactions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 671−678.
(17) Satoh, H.; Funatsu, K.; Takano, K.; Nakata, T. Classification and Prediction of Reagents' Roles by FRAU System with Self-Organizing Neural Network Model. *Bull. Chem. Soc. Jpn.* **2000**, *73*, 1955−1965.
(18) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of 2 Different Neural-Network Types for the Prediction of Chemical-Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.
(19) Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (COMSA): a Novel Tool for Molecular Design. *Comput. Chem.* **2000**, *24*, 615−625.
(20) Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J. The Representation of Molecular Electrostatic Potentials by Topological Feature Maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.
(21) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19−64.
(22) Kohonen, T. *Self-Organization and Associative Memory*; Springer: Berlin, 1988.
(23) PETRA can be tested on the website http://www2.chemie.uni-erlangen.de and is available from Molecular Networks GmbH, http://www.mol-net.de.

(24) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds; Springer-Verlag: Heidelberg, Germany, 1988; pp 119−138.
(25) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(26) Aires-de-Sousa, J. JATOON: Java Tools for Neural Networks. *Chemom. Intell. Lab. Syst*. **2002**, *61*, 167−173.
(27) The JATOON applets are available at http://www.dq.fct.unl.pt/staff/jas/jatoon.
(28) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(29) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1947−1958.
(30) *R: A language and environment for statistical computing.* R Development Core Team, R Foundation for Statistical Computing: Vienna, Austria, 2004. ISBN 3-900051-07-0; URL http://www.R-project.org.
(31) Fortran original by Leo Breiman and Adele Cutler; R port by Andy Liaw and Matthew Wiener. http://www.stat.berkeley.edu/users/breiman/ (2004).

CI0502707