

Machine learning of chemical reactivity from databases of organic reactions

Gonalo V. S. M. Carrera · Sunil Gupta ·
Joo Aires-de-Sousa

Received: 27 November 2008 / Accepted: 18 April 2009 / Published online: 26 May 2009
© Springer Science+Business Media B.V. 2009

Abstract Databases of chemical reactions contain knowledge about the reactivity of specific reagents. Although information is in general only explicitly available for compounds reported to react, it is possible to derive information about substructures that do not react in the reported reactions. Both types of information (positive and negative) can be used to train machine learning techniques to predict if a compound reacts or not with a specific reagent. The whole process was implemented with two databases of reactions, one involving BuNH₂ as the reagent, and the other NaCNBH₃. Negative information was derived using MOLMAP molecular descriptors, and classification models were developed with Random Forests also based on MOLMAP descriptors. MOLMAP descriptors were based exclusively on calculated physicochemical features of molecules. Correct predictions were achieved for ~90% of independent test sets. While NaCNBH₃ is a selective reducing reagent widely used in organic synthesis, BuNH₂ is a nucleophile that mimics the reactivity of the lysine side chain (involved in an initiating step of the mechanism leading to skin sensitization).

Keywords MOLMAP · Chemical reactivity · Databases · Machine learning · Electrophilicity

Abbreviations

MOLMAP	MOlecular maps of atom-level properties
BuNH ₂	Butylamine
RF	Random forest
VOC	Volatile organic compounds
QSAR	Quantitative structure activity relationship
OOB	Out of bag
SVM	Support vector machines
ROC	Receiver operating characteristic
SOM	Self organizing maps
HTS	High-throughput screening

Introduction

Chemoinformatics approaches that learn from available experimental data to make rapid estimations of chemical reactivity are currently sought for various applications in different fields. Chemical reactivity is involved in toxicological mechanisms responsible for skin sensitization, [1] mutagenicity, [2] or adverse side effects of drugs [3]. Prediction of reactivity is needed in pharmaceutical R&D innovation processes, or for the prioritization of experimental tests in risk assessment of chemicals, namely in relation with the EU REACH [4] legislation. Furthermore, the legislative trend for the abolition of animal testing of cosmetic products [5] is demanding alternative evaluation procedures [6, 11]. For the assessment of skin sensitization, in vitro reactivity tests, [7–10] as well as QSARs have been proposed. In silico methodologies are of interest also for “Integrated Testing Strategies” that combine different types of data and information, e.g. predictions or results obtained from several single tests, in the decision-making process [10]. In the area of eco-toxicology,

Electronic supplementary material The online version of this article (doi:10.1007/s10822-009-9275-2) contains supplementary material, which is available to authorized users.

G. V. S. M. Carrera · S. Gupta · J. Aires-de-Sousa (✉)
REQUIMTE, CQFB, Departamento de Qumica, Faculdade
de Cincias e Tecnologia, Universidade Nova de Lisboa,
2829-516 Caparica, Portugal
e-mail: jas@fct.unl.pt

modeling the environmental fate of compounds requires the prediction of atmospheric reactions involving radicals, which are responsible for the degradation of pollutants [12].

As compounds react differently with different reagents, chemical reactivity must be defined in relation to a specific reagent. Data sets of compounds and the corresponding reactivity towards a common reagent can be used to establish structure-reactivity relationships in a similar way to QSAR approaches, based on machine learning or statistical techniques. However, most QSAR studies have focused on properties resulting from the global molecular structure, such as the affinity to a biological target, or aqueous solubility. In contrast, chemical reactivity mostly depends on local properties, typically functional groups and their near environment. Furthermore, data on chemical reactivity are available in very different frameworks of databases, and with varying degrees of fuzziness, ranging from databases of rate constants measured under specific reaction conditions, to databases of chemical reactions (with or without assignment of reaction centers), to lists of compounds known to be stable or unstable under more or less defined conditions. Different approaches are required for different scenarios, as illustrated by the following two examples. In some applications the occurring reaction sites and transformations are known, in others information is only available whether a compound “reacts or not”. Data sets may include series of compounds with a common reaction center, or may include different reaction centers, unknown reaction centers, or compounds with more than one reaction center.

Regressions have been derived to predict rate constants from molecular descriptors of compounds. Such models are typically built with data sets of compounds with common reaction centers (where the reaction is known to take place), and for which the rate constants are available [13–15]. The selected molecular descriptors have included quantum chemical descriptors, empirical physicochemical descriptors, as well as descriptors directly calculated from the molecular connectivity. Descriptors have been defined for the whole molecule or for specific atoms of the reaction center. Examples of reported studies include the estimation of ester hydrolysis rate constants with quantum descriptors, [13, 14] and the estimation of gas-phase homolysis rate constants with molecular descriptors calculated by CODESSA [15]. A much studied topic is the estimation of rate constants of degradation reactions with OH and NO₃ radicals, and ozone, as these mainly control the tropospheric lifetime of most organic chemicals, including volatile organic compounds (VOCs) [12, 16, 17]. Models have been established with fragment/group contributions, [18] multi-linear regressions, [19] non-linear equations, [20] as well as neural networks [21].

Other studies focused on classification tasks rather than regression. Hajduk and coworkers [22] developed a filtering tool based on substructural fragments to classify compounds as reactive or nonreactive towards the human La antigen, a reactivity probe in HTS of protein targets. Satoh et al. [23, 24] devised the FRAU method to numerically characterize a field around an atom in a molecule, which is based on electrostatic and steric interactions with a pseudoreactant. These codes were applied to the classification of atoms by self-organizing maps in order to classify the roles of the corresponding reagents in reactions.

Gasteiger et al. [25] studied empirical physicochemical properties of covalent bonds for their application in models of chemical reactivity. These included properties for the bond atoms, such as electronegativities, atomic charges, or polarizabilities, as well as features of the bonds, such as resonance stabilization after bond breaking, bond dissociation energy, or differences between atomic properties of bond atoms (e.g. atomic charges).

In principle, reactivity models based on physicochemical properties are in a better position to capture relationships of wider applicability. In order to encode the physicochemical features of bonds for an entire molecule, and at the same time having a fixed-length alignment-free representation of the molecule, independent of its size, we proposed to map all the bonds of a molecule into a 2D self-organizing map—a *MOLMAP* (MOlecular Map of Atom-level Properties) [26, 27]. MOLMAPs encode the types of bonds available in a molecule. In a chemical reaction, the difference between the MOLMAPs of the products and the MOLMAPs of the reactants has been used as a descriptor of the reaction, [28] and has been applied to the classification of large data sets of enzymatic reactions [29, 30].

In this article we report machine learning studies with a database of >900 reactions involving butylamine (BuNH₂) to predict whether a compound reacts or not with BuNH₂. As only positive data was explicitly available, i.e. the database only includes reactions, not compounds that do not react, a strategy had to be developed to extract negative data. Negative data were derived from parts of the reactants that were unchanged in the reactions.

Besides exploring the possibility of applying machine learning techniques to extract knowledge from databases of organic reactions in order to predict chemical reactivity, which is the aim of this paper, butylamine was chosen as it is a model nucleophile of the side chain of lysine. This kind of reactivity is involved in nucleophilic additions of skin proteins to chemicals, triggering skin sensitization that may possibly result in allergic contact dermatitis. In vitro studies have linked reactivity towards amino acids with potency of skin sensitization [7–9]. Thus the prediction of chemical reactivity towards lysine side chains is helpful to assess the potential for a chemical to be a skin sensitizer.

The model here generated is based on a very inclusive data set (with diverse reaction conditions, and no account of reaction times or yields even if they were available in the original database) and is therefore not directly applicable to in vitro reactivity tests with peptides, which are performed under well-defined conditions. Even though, we compared, as an exercise, the predictions of our model with reported experimental data for reactivity towards a lysine peptide. Clearly, as reactivity depends on experimental conditions, a model trained with data obtained under unspecified conditions can only be expected to learn reactivity features revealed in most available cases.

The devised methodology was applied with a database of reactions involving sodium cyanoborohydride (NaCNBH_3), which is a selective reducing reagent widely used in organic chemistry. It was chosen here to test the ability of MOLMAPs to differentiate between similar functional groups.

With this strategy we simulated (under limitations here discussed) a situation where reactivity data is available in the form of a list of compounds that react towards a specific reagent, and a list of compounds that do not react with the same reagent. No information was used about the reaction site or the type of reaction that could occur, and only physicochemical features were used to process reaction data.

Methodology

Data sets

A database of reactions involving butylamine was extracted from the SPRESI database (Infochem GmbH, Munich, Germany). From these, a data set of 971 reactions was selected with the following criteria: (1) reactions involving only one reactant (other than butylamine) and only one product, (2) reactions stoichiometrically balanced within a tolerance of +6 atoms of difference between reactants and product, (3) reactions involving no catalyst other than butylamine, and (4) reactions involving reactants and products accepted by the PETRA program [31]. PETRA was used to calculate the empirical physicochemical features of bonds, for the generation of the MOLMAP descriptors (see below). The data set was randomly divided into a training set with 922 reactions and a test set with 49 reactions assuring that the test set covered the different types of reactive functional groups in the database. For the training set, negative information (compounds that do not react) was derived as MOLMAPs of pseudocompounds (see below). From the 49 reactants in the test set, a corresponding set of 49 molecules that do not react were obtained by manually removing the reactive functional groups. This test set was further partitioned in half to get a validation set (used to monitor the optimization of the

models) and an independent prediction set. The 922 reactive compounds of the training set have an average molecular weight of 250.35 (standard deviation of 115.15). They usually include more than one functional group, 762 of them have at least one oxygen atom (average of 2.24 O-atoms/molecule), and 606 have at least one nitrogen atom (average of 1.46 N-atoms/molecule).

A second database of reactions was extracted from the SPRESI database involving NaCNBH_3 as a reactant, and further filtered using the same criteria as for the database of BuNH_2 reactions. In this case differences between the number of atoms in the reactants (excluding NaCNBH_3) and in the products between -6 and $+8$ were accepted. From the obtained data set of 364 reactions a random set of 319 reactants was extracted as the training set. The data set with the remaining 45 reactants was further partitioned into a validation set (22 reactants) and an independent prediction set (23 reactants). For the training set, negative information (compounds that do not react) was derived as MOLMAPs of pseudocompounds (see below). The negative information for the validation and prediction sets was generated as for the BuNH_2 dataset, by manually removing the reactive functional groups. Two more reactants (with functional groups known to react but not available otherwise in the test set, positive objects) were added to the validation set as well as to the independent test set. Therefore the training set consists of 319 positive and 319 negative objects, the validation set has 24 positive and 22 negative objects, and the independent test set has 25 positive and 23 negative objects. The 319 reactive compounds of the training set have an average molecular weight of 279.01 (standard deviation of 124.78). They usually include more than one functional group, 279 of them have at least one oxygen atom (average of 2.73 O-atoms/molecule), and 273 have at least one nitrogen atom (average of 1.41 N-atoms/molecule).

Generation of MOLMAP molecular descriptors

MOLMAP molecular descriptors [26] are based on a Kohonen self-organizing map (SOM) trained with a diversity of covalent bonds taken from a set of structures.

A Kohonen SOM [32] is an unsupervised method that projects multidimensional objects (in this case bonds) into a 2D surface (a map of neurons) which can reveal similarities between objects, mapped into the same or neighbor neurons. In this study SOMs were trained with 4980 bonds randomly chosen from the reactants of the database of BuNH_2 reactions, and represented by physicochemical features. Training was performed by using a linear decreasing triangular scaling function with an initial learning rate of 0.1 and an initial learning span of half the size of the map. The weights were initialized with random numbers that

were calculated using as parameters the mean and standard deviation of the corresponding variables in the input data set. For the selection of the winning neuron, the minimum Euclidean distance between the input vector and neurons was used. The training was typically performed over 100 cycles, with the learning span and the learning rate linearly decreasing until zero. SOMs were implemented throughout this study with an in-house developed Java application.

In order to obtain empirical physicochemical features of atoms and bonds, the molecules were submitted to the PETRA program (version 4). In the experiments here reported, bonds (A–B) were represented by the following physicochemical features: charge for atoms A and B (sigma charge, pi charge, and total charge), sigma electronegativity for atoms A and B, bond polarity, and resonance stabilization of the charges produced in heterolytic bond dissociation (considered in both directions). As descriptors for a bond A–B are different if the bond is considered from A to B, or from B to A, each bond was here always considered from the atom with higher total atomic charge to the atom with lower charge. Bond features were z-normalized. To focus on substructures around functional groups, only bonds were considered that include (or are at a one bond distance from) a heteroatom or an atom belonging to a pi system.

Based on the SOM previously trained with a diversity of bonds, the MOLMAP descriptor for a molecule is the pattern of neurons activated by its bonds. It can be interpreted as a representation of the available bonds in the molecule. 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 columns (MOLMAPs of type 1:0). In order to account for the relationship between similarity of bonds and proximity in the map, alternative MOLMAP descriptors were generated (MOLMAPs of type 10:1), in which a value of 10 was assigned to each neuron multiplied by the number of times it was activated by a bond, and a value of one was added to each neuron multiplied by the number of times a neighbor was activated by a bond.

With this approach, bonds are clustered and differentiated on the basis of their physicochemical parameters, and MOLMAPs completely rely on this mapping. So the successes and failures of the whole approach are linked to the quality of the physicochemical parameters and the mapping.

Generation of MOLMAP descriptors of pseudocompounds that do not react

In order to train a classifier to predict if a compound reacts or not, both compounds known to react (positives) and compounds known to not react (negatives) are needed.

However, the database of reactions involving a specific reagent does not include failed reactions, i.e. there is no explicit information about compounds that do not react with the specific reagent. In the absence of such information we derived negative information from (descriptors of) the substructures of the reactants inferred not to include the reaction centers (for a similar idea see reference [24]). In most of the reactions, reactants are expected to have only one possible site for reaction with the specific reagent (BuNH_2 or NaCNBH_3). The regions of the molecule not involved in the reaction can thus be generally used as negative examples.

MOLMAPs of *pseudocompounds* that do not react were directly derived from the MOLMAPs of the products and reactants of the reactions in the databases. In a reaction, the comparison of the MOLMAP of the product with the MOLMAP of the reactant reveals MOLMAP components with the same or higher values in the product than in the reactants—these correspond to bonds with mostly unchanged features that can generally be considered non-reacting bonds. A MOLMAP consisting of the components of the reactant that remain over the reaction is therefore used as a MOLMAP of a *pseudocompound* that does not react—a negative example required for the training of the classifier.

Random forests for the classification of reactivity

A Random Forest (RF) 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 [33, 34]. Prediction is made by majority vote of the individual trees. Additionally, performance is assessed with the prediction error for the objects left out in the bootstrap procedure—out-of-bag estimation (OOB). The method also quantifies the importance of each variable, and provides a proximity measure between objects based on the number of times they reach the same terminal nodes of the forest trees.

The RF is a high-dimensional nonparametric method that works well on large numbers of variables. It was chosen as the main machine learning algorithm for this study since the data set incorporates a large number of variables but those relevant for reactivity are expected to be few. Because reactivity is expected to result from the presence of certain types of bonds (encoded as components of the MOLMAP descriptor), a method based on recursive partitioning seemed appropriate. Additionally, RFs are generally a more accurate classifier than individual classification trees.

In this study, Random Forests were grown with the R program version 2.0.1 [35] using the random Forest library

[36]. Forests with 1000 trees were used, typically with default values for other parameters. RF were trained with MOLMAPs of reactants (positive cases) and MOLMAPs of pseudocompounds that do not react (negative cases) to predict whether a compound reacts or not with a specific reagent (BuNH_2 or NaCNBH_3).

Results and discussion

Prediction of reactivity with BuNH_2

Random Forests with 1000 trees and 25 random variables made available at each node (*mtry*) were trained with MOLMAPs of reactants (classified as “reactive”) and MOLMAPs of pseudocompounds that do not react (classified as “non-reactive”) for the task of discriminating between the two classes. Different sizes of MOLMAPs were tried, as well as two different types (10:1 and 1:0, see Methodology). The ability of the models to predict reactivity was assessed by the out-of-bag (OOB) estimation for the training set, and by the accuracy of predictions for the validation set. The results are presented in Table 1. The RF trained with MOLMAPs smaller than 20×20 exhibited a high percentage of correct predictions in the OOB estimation but poor performance with the validation set. This was particularly evident for MOLMAPs of type 10:1. It can be explained by the fact that small MOLMAPs have a higher proportion of components changed from the reactants to the products, causing in a high proportion of null components in the derived MOLMAPs of non-reactive pseudocompounds. Therefore, the MOLMAPs of the two classes (reactive and non-reactive) become artificially trivial to distinguish (leading to good OOB predictions) but having no significance as observed with the predictions for real molecules (validation set). Optimization of the models with the validation set is thus recommended.

In alternative to RF, a single decision tree was also trained with MOLMAPs of type 1:0 and size 29×29 . This could only correctly predict 65% of the validation set, but

provided useful insights about the inferred knowledge. (It is expected that rules inferred by RFs are similar to the rules of the single decision tree.) Inspection of the tree revealed that some rules were based on MOLMAP components associated to bonds of BuNH_2 . The breaking or changing of such bonds when BuNH_2 covalently binds a molecule make the corresponding MOLMAP components to become null in pseudoreactants that do not react, which leads to the perception of amines (where such bonds exist) as reactive compounds. This observation was confirmed by training new RF with the same data set but excluding from the list of variables the components corresponding to the bonds of BuNH_2 —accuracy of predictions was improved to 92% for the validation set and 76% in OOB (not shown in the Table).

In the same direction, MOLMAPs (type 1:0) of non-reactive pseudocompounds were generated by comparing exclusively the product with the reactant, not including BuNH_2 as a reactant. The negative data thus obtained, and the MOLMAPs of reactive compounds, were used to train new Random Forests (Table 2).

This modification also improved the percentage of correct predictions to 89.80% (validation set) in MOLMAPs of size 29×29 and 20×20 . These were the sizes of MOLMAP yielding the best predictions. Considering both sizes, the RF *mtry* parameter was optimized in the range between 25 and 300, which lead to 93.88% of correct predictions for the validation set with the optimized value of *mtry* = 50 (obtained with MOLMAPs of size 20×20).

An inspection of the RF-assigned relative importance of descriptors for this model revealed S–C bonds ($-\text{SCH}_3$ group) and C–Cl bonds associated with the most important descriptors. SCH_3 and Cl are indeed common leaving groups in this dataset of reactants. Among the ten most important are descriptors associated with the C–O bond of epoxides, C=O bond of carboxylic acids and esters, C–O bond of alkoxide group in esters, C–C=O bond of ketones, C–C bond connecting an electro-withdrawing group (nitrile and nitro) and a C=C bond, and C=O bond of aldehydes and ketones. All these types of bonds commonly react

Table 1 Prediction of reaction with BuNH_2 by Random Forests trained with different MOLMAPs

Size of MOLMAP	MOLMAP type 10:1		MOLMAP type 1:0	
	% OOB correct predictions	% validation set correct predictions	% OOB correct predictions	% validation set correct predictions
29×29	85.52	75.51	75.81	81.63
25×25	87.69	65.31	74.46	77.55
20×20	91.76	69.39	75.43	81.63
15×15	96.10	61.22	76.63	71.42
10×10	97.40	55.10	82.38	75.51
7×7	98.26	55.10	82.92	71.42

Table 2 Prediction of reaction with BuNH₂ by random forests trained with MOLMAPs of type 1:0 (non-reactive pseudocompounds generated without including BuNH₂ as reactant)

Size of MOLMAP	% OOB correct predictions	% validation set correct predictions
29 × 29	74.30	89.80
25 × 25	74.46	85.71
20 × 20	74.08	89.80
15 × 15	74.95	83.67
10 × 10	79.12	67.34
7 × 7	73.81	85.71

(or at least significantly influence the outcome of the reaction) in the presence of a nucleophile as butylamine.

Additionally, a decision tree was trained with the same data, resulting in 13 rules and yielding 79.59% of correct assignments for the validation set. Inspection of the rules revealed that they were all positive rules, i.e. prediction of the *reactive* class was inferred from the *presence* of certain features. The types of bonds associated with the rules were essentially the same as for the most important descriptors selected by RF model. The decision tree additionally included descriptors related to C–Cl bonds of acyl chlorides and C=S bonds of isothiocyanates.

The selected RF model was further tested with the independent prediction set of 49 compounds, not used so far, achieving 87.76% of correct predictions with sensitivity = 0.79 and specificity = 0.96. The six wrongly classified objects include five false negatives and one false positive. Wrong predictions can be explained by a lack of similar examples in the training set (revealed by a low similarity measure to the most similar object in the training set, as calculated by the RF in-bound procedure). Four of the problematic predictions have assigned probabilities not far from 0.5. The other two are false negatives, one is a compound with a unique functionality absent from the training set, the other has a reacting substructure that does not react in the most similar object of the training set (due to reaction selectivity). The association between the assigned probabilities and the experimental classes is illustrated in Fig. 1.

With sensitivity and specificity values obtained at different thresholds of probability, considering the prediction set of 49 compounds, a ROC curve was generated in order to assure that the increment in the rate of true positives is not related to the increment of the total number of positive predictions independently of being true or false (Fig. 2). This measure of uncorrelation is the area between the ROC curve and the line of no discrimination. A value of 0.93 was obtained.

The model was also validated by y-randomization. The classification of the objects in the training set was

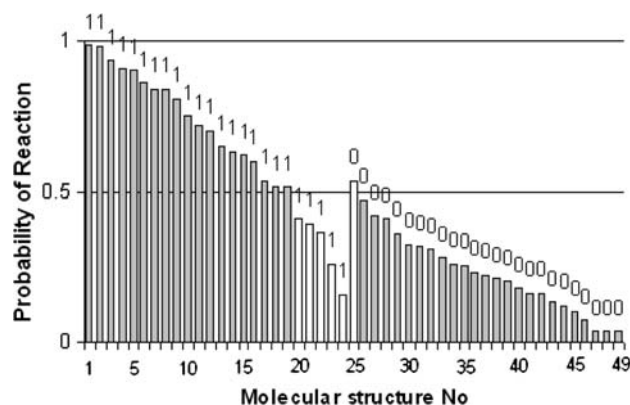


Fig. 1 Association between the experimental reactivity class (1—reactive, 0—non-reactive) and the RF-assigned probability of reaction with BuNH₂ for the 49 objects of the independent prediction set. Correct predictions in gray bars, wrong predictions in white bars

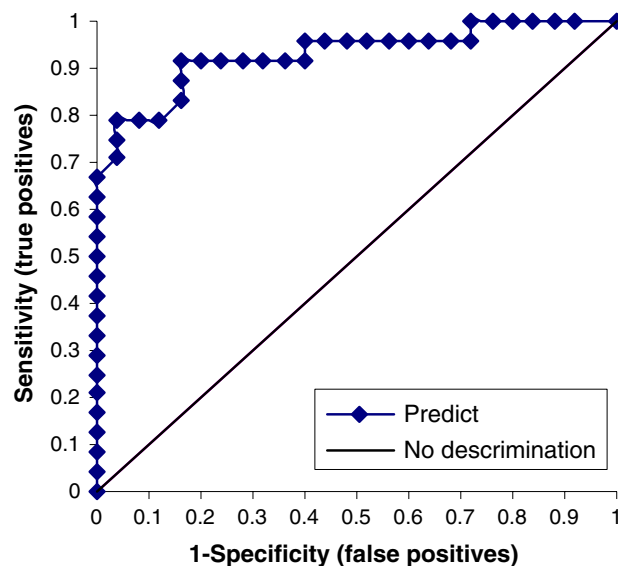


Fig. 2 ROC obtained with the RF results for the 49 compounds of the prediction set

randomly reassigned, keeping the 1:1 proportion of “reactive” and “non-reactive” classifications. A RF trained with the modified training set gave an OOB estimation error of 50%, and correct predictions for the validation and prediction sets of 59 and 45%, respectively.

Support vector machines (SVM) were also tried with the same data set, as an alternative machine learning technique, but the performance was inferior to Random Forests. Correct predictions for the independent prediction set only attained 82%.

As an exercise, the final RF model was applied to a new set of 75 compounds, and predictions were compared with in vitro reported results for the reaction with a peptide containing lysine [8, 9]. The reactivity was measured as the percentage of the peptide depletion caused by the

compound. Although the papers of Gerberick et al. included some more compounds, a few were filtered out—duplicates of the training set, mixtures of compounds and chemicals with no available reactivity information. The results are in Table S1 of the Supplementary Material.

Almost 60% of the compounds with reported reactivity <10% are predicted as nonreactive, and 58% of compounds with reactivity >10% are predicted to react. Most false positives can be easily explained. Carboxylic acids (e.g. octanoic acid, 4-hydroxybenzoic acid and lactic acid) are predicted to react although they almost do not cause peptide depletion. They are predicted to react because carboxylic acids in the training set such as octadecanoic acid, benzoic acid and 2,2-dimethylpropionic acid are described to react with BuNH₂. The incongruence is most probably related to different experimental conditions in the specific peptide test and in the reactions of the training database. Also aldehydes and ketones usually react in database entries, but non-activated ketones do not react in the peptide assay, and a number of aldehydes are reported as unreactive or only weakly reactive (e.g. α -hexylcinnamaldehyde, hydroxycitronellal, and ethyl vanillin). The different reactivity of α -hexylcinnamaldehyde, a non-reactive α,β -unsaturated aldehyde with a long alkyl substituent at the α position, and cinnamic aldehyde (43.2% depletion in the assay) cannot be fully perceived by the MOLMAP descriptors and the information in the database (although the probability of reaction assigned by the RF to the latter is substantially higher than to the former). Nonanoyl chloride and 1-bromobutane are predicted as reactive because acyl chlorides and haloalkanes generally react in the database, although <2% peptide depletion is reported for those two cases.

Activated esters also appear as problematic. They are usually predicted as non-reactive, although some exhibit in vitro reactivity above 10%. In the training set, the closest objects to lauryl gallate, 2-hydroxyethyl acrylate, diethyl maleate, and ethyleneglycol dimethacrylate are pseudocompounds derived from aromatic or α,β -unsaturated esters, which include other functional groups described to react leaving the unsaturated ester functional group intact. This selectivity problem is inherent to the extraction of negative information from non-reacting substructures of reactive compounds. Another explanation for poor results with esters may be the exclusion of reactions from the database that are stoichiometrically unbalanced (outside a tolerance of six atoms)—several aminolysis of esters with alcohol leaving groups larger than methanol were removed (these leaving groups are generally not listed as products in the database).

p-Benzoquinone and 1,4-hydroquinone are wrongly predicted to not react. 1,4-hydroquinone is predicted as non-reactive because phenols are not generally perceived

as reactive. In fact, it is probably air oxidized prior to reaction with lysine, and the RF model does not account for any kind of activation (metabolic or non-metabolic). The surprising prediction of benzoquinone as non-reactive probably results from a lack of Michael additions in the database and even the presence of at least one example of a non-reacting substructure with an α,β -unsaturated ketone.

Prediction of reactivity with NaCNBH₃

Differently from reactions with BuNH₂, reactions with NaCNBH₃ do not result in the addition of atoms other than hydrogen. Negative information corresponding to pseudocompounds that do not react was therefore always obtained from the comparison of the MOLMAP of the product with the MOLMAP of the reactant.

Random Forests trained to predict whether a compound reacts or not with NaCNBH₃ were trained similarly to the study with BuNH₂. Results are displayed in Table 3 for the OOB estimation, and for the validation set comprising 46 objects. The best performance was observed with the MOLMAPs of type 1:0 and size 15 \times 15. The model could be further improved by changing the value of *mtry* to 75. This model was able to correctly predict 100% of the validation set.

Application of the best model to the independent prediction set of 48 objects yielded 94% of correct predictions, sensitivity = 0.88 and specificity = 1.00. The area under the ROC curve is 0.991. The three wrong predictions correspond to false negatives, all three with assigned probabilities of reaction above 0.4. Their most similar objects in the training set are pseudocompounds (non-reactive) corresponding to very similar non-reactive substructures both in terms of skeleton and functional groups. In two of them, reaction of a specific hydroxyl group is (wrongly) not predicted. In their closest reactions of the database, a similar hydroxyl group reacts but the same

Table 3 Prediction of reaction with NaCNBH₃ by random forests trained with different MOLMAPs

Size of MOLMAP	MOLMAP type 10:1		MOLMAP type 1:0	
	% correct predictions		% correct predictions	
	OOB	Validation set	OOB	Validation set
29 \times 29	91.70	91.30	87.93	93.48
25 \times 25	92.16	89.13	87.62	91.30
20 \times 20	92.63	82.61	86.68	89.13
15 \times 15	96.39	80.43	89.03	95.65
10 \times 10	97.81	65.22	89.18	84.78
7 \times 7	98.75	60.87	91.69	82.61

reaction also reduces a C=C bond, which is not present in the query compounds.

Concerning the similarity between objects in the training set and in the test sets, it was found that there were only three objects in the validation set, and three objects in the prediction set with the same MOLMAP as objects in the training set (although the corresponding molecules were not the same).

The model was also validated by y-randomization as described for the study with BuNH₂, obtaining an OOB estimation error of 53%, and correct predictions for 52 and 46% of the validation and prediction sets, respectively.

Among the ten most important MOLMAP descriptors listed by the RF model were those associated with N=C bonds of imines, C=S bonds of dithioalkylic esters, C=C bonds α,β conjugated to electron-withdrawing groups like carbonyl, C–C bonds directly connected to carbonyl or to α,β -unsaturated carbonylic groups, C–H bonds directly connected to α,β -unsaturated carbonylic groups, C=O and C–C=O bonds of aldehydes and ketones.

A decision tree based on the same data set also performed well. It could correctly predict 85% of the prediction set. The inferred 11 rules were based on MOLMAP descriptors associated with essentially the same types of

Table 4 List of compounds derived from literature information concerning reactivity with sodium cyanoborohydride with assigned experimental/predicted class (1—reactive, 0—non reactive)





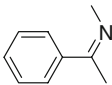



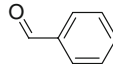


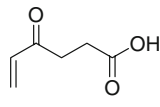
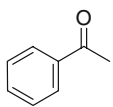


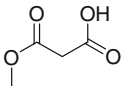
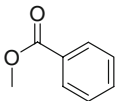
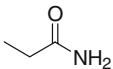


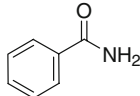


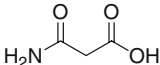
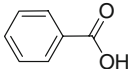
 Isopropylidene-methyl-amine 1/1	 Isopropylideneamine 1/1	 Methyl-(1-methyl-allylidene)-amine 1/1
 Ethylidene-methyl-amine 1/1	 Methyl-(1-phenyl-ethylidene)-amine 1/1	 Acetaldehyde 1/1
 Propenal 1/1	 Formaldehyde 1/1	 Benzaldehyde 1/1
 Propan-2-one 1/1	 But-3-en-2-one 1/1	 4-Oxo-hex-5-enoic acid 1/1
 1-Phenyl-ethanone 1/1	 Propionic acid methyl ester 0/0	 Acrylic acid methyl ester 0/0

Table 4 continued

 <p>Malonic acid monomethyl ester 0/0</p>	 <p>Benzoic acid methyl ester 0/0</p>	 <p>Propionamide 0/0</p>
 <p>N,N-Dimethyl-propionamide 0/0</p>	 <p>N-Methyl-acrylamide 0/0</p>	 <p>Benzamide 0/0</p>
 <p>Acetic acid 0/0</p>	 <p>Acrylic acid 0/0</p>	 <p>Malonamic acid 0/0</p>
 <p>Benzoic acid 0/0</p>		

bonds as for the most important descriptors in the RF model. Additionally, descriptors associated with the C–O bonds of epoxides were selected.

The best RF model was further validated with a set of 25 compounds extracted from the literature [37] with simple skeletons and encompassing different functional groups, to probe the ability to predict the chemoselectivity of NaC–NBH₃ (Table 4). The percentage of correct assignments was 100%. The results show that the model, which was built exclusively from physicochemical descriptors, can distinguish between the reactive functionalities imine, ketone, or aldehyde, from non-reactive functionalities such as ester, carboxylic acid and amide.

In both applications (BuNH₂ and NaC–NBH₃) information about the reaction centers was not used (although reaction centers are assigned in the SPRESI databases) in order to simulate a situation where only a list of compounds that react with a given reagent, and a list of compounds that do not react, are known. In such a practical situation the data set would in general not include pairs of reactive and non-reactive compounds only differing in the

reaction center, which would make learning harder. Under defined experimental conditions, variation of structural features in compounds with a common (generally reactive) functional group modulates their reactivity, and may render some of them unreactive. That too would make learning harder. On the other hand, experimental negative data (compounds experimentally known not to react) would avoid the inclusion of false negatives, in contrast with the approach here explored (due to selectivity issues in compounds with multiple reactive sites but only one shown in the database entry).

The information available in the SPRESI database about reaction centers could have been used to derive regions of the reactants that do not react. However, it would have required a decision about the accepted distance from the reaction center for a bond to belong to a “non-reactive substructure”. Also, the use of MOLMAPs encoding local physicochemical properties of molecules allows for the recognition of patterns of bonds responsible for reaction, in addition to the recognition of single bonds with certain patterns of properties.

Conclusions

A study on chemical reactivity could be performed exclusively based on empirical physicochemical properties calculated for atoms and bonds in molecules.

MOLMAP molecular descriptors enabled to derive negative information about reactivity towards two specific reagents (BuNH_2 and NaCNBH_3) from databases of chemical reactions involving those reagents. From this information, together with MOLMAPs of reactive compounds (reactants in the database), Random Forests could extract knowledge on the reactivity of compounds towards the specific reagents, which was validated with independent prediction sets. Negative information derived from MOLMAPs consists of bonds that are not influenced by the changes operated by the reaction. A drawback of deriving negative information from reactions that did occur is the inclusion of false negatives when reactants have more than one reaction site.

The method requires the training data to be extracted from reactions that operate changes of properties on the bonds at the reaction center (even if explicit information about the reaction center is not used). Additionally, reactions should be stoichiometrically balanced.

The results demonstrated the MOLMAP/machine learning approach to automatically extracting reactivity insights from databases of reactions. In its current stage, however, the model for BuNH_2 cannot be applied to solve risk assessment problems related to skin sensitization. For that matter, more realistic models are expected to become possible with this approach as larger databases emerge with experimental reactivity data towards peptides measured under standard conditions that can be used for training.

Acknowledgments G.C. and S.G. acknowledge Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for financial support under grants SFRH/BD/18354/2004 and SFRH/BPD/14475/2003. Molecular Networks GmbH (Erlangen, Germany) and Infochem (Munich, Germany) are acknowledged for access to the PETRA program and to subsets of chemical reactions from the SPRESI database, respectively.

References

1. Aptula AO, Patlewicz G, Roberts DW (2005) *Chem Res Toxicol* 18:1420. doi:10.1021/tx050075m
2. Benigni R (2005) *Chem Rev* 105:1767. doi:10.1021/cr030049y
3. Metz JT, Huth JR, Hajduk PJ (2007) *J Comput Aided Mol Des* 21:139. doi:10.1007/s10822-007-9109-z
4. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm
5. Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC. OJ L066, 26–35, 11 March 2003
6. Lilienblum W, Dekant W, Foth H, Gebel T, Hengstler JG, Kahl R, Kramer P-J, Schweinfurth H, Wollin K-M (2008) *Arch Toxicol* 82:211. doi:10.1007/s00204-008-0279-9
7. Aptula AO, Patlewicz G, Roberts DW, Schultz TW (2006) *Toxicol In Vitro* 20:239. doi:10.1016/j.tiv.2005.07.003
8. Gerberick GF, Vassallo JD, Bailey RE, Chaney JG, Morrall SW, Lepoittevin J-P (2004) *Toxicol Sci* 81:332. doi:10.1093/toxsci/kfh213
9. Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin J-P (2007) *Toxicol Sci* 97:427. doi:10.1093/toxsci/kfm064
10. Natsch A, Emter R, Ellis G (2009) *Toxicol Sci* 107:106. doi:10.1093/toxsci/kfn204
11. Patlewicz G, Aptula AO, Roberts DW, Uriarte E (2008) *QSAR Comb Sci* 27:60. doi:10.1002/qsar.200710067
12. Gramatica P, Pilutti P, Papa E (2004) *Atmos Environ* 38:6167. doi:10.1016/j.atmosenv.2004.07.026
13. Chaudry UA, Popelier PLA (2003) *J Phys Chem A* 107:4578. doi:10.1021/jp034272a
14. Zhang H, Qu X, Ando H (2005) *J Mol Struct THEOCHEM* 725:31. doi:10.1016/j.theochem.2005.02.086
15. Hiob R, Karelson M (2000) *J Chem Inf Comput Sci* 40:1062. doi:10.1021/ci0004457
16. Meylan WM, Howard PH (2003) *Environ Toxicol Chem* 22:1724. doi:10.1897/01-275
17. Gramatica P, Consonni V, Todeschini R (1999) *Chemosphere* 38:1371. doi:10.1016/S0045-6535(98)00539-6
18. Atkinson R (1998) *Environ Toxicol Chem* 7:435. doi:10.1897/1552-8618(1998)7[435:EOGHRR]2.0.CO;2
19. Gramatica P, Pilutti P, Papa E (2004) *J Chem Inf Comput Sci* 44:1794
20. Klamt A (1993) *Chemosphere* 26:1273. doi:10.1016/0045-6535(93)90181-4
21. Fatemi MH (2006) *Anal Chim Acta* 556:355. doi:10.1016/j.aca.2005.09.033
22. Huth JR, Mendoza R, Olejniczak ET, Johnson RW, Cothron DA, Liu Y, Lerner CG, Chen J, Hajduk PJ (2005) *J Am Chem Soc* 127:217
23. Satoh H, Itono S, Funatsu K, Takano K, Nakata TA (1999) *J Chem Inf Comput Sci* 39:671. doi:10.1021/ci9801567
24. Satoh H, Funatsu K, Takano K, Nakata T (2000) *Bull Chem Soc Jpn* 73:1955. doi:10.1246/bcsj.73.1955
25. Simon V, Gasteiger J, Zupan J (1993) *J Am Chem Soc* 115:9148. doi:10.1021/ja00073a034
26. Gupta S, Mathew S, Abreu PM, Aires-de-Sousa J (2006) *Bioorg Med Chem* 14:1199. doi:10.1016/j.bmc.2005.09.047
27. Zhang Q, Aires-de-Sousa J (2007) *J Chem Inf Model* 47:1. doi:10.1021/ci050520j
28. Zhang Q-Y, Aires-de-Sousa J (2005) *J Chem Inf Model* 45:1775. doi:10.1021/ci0502707
29. Latino DARS, Aires-de-Sousa J (2006) *Angew Chem Int Ed* 45:2066. doi:10.1002/anie.200503833
30. Latino DARS, Zhang Q-Y, Aires-de-Sousa J (2008) *Bioinformatics* 24:2236. doi:10.1093/bioinformatics/btn405
31. <http://www2.chemie.uni-erlangen.de/software/petra/>
32. Kohonen T (1998) *Self-Organization and Associative Memory*. Springer, Berlin
33. Breiman L (2001) *Mach Learn* 45:5. doi:10.1023/A:1010933404324

34. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BPJ (2003) Chem Inf Comput Sci 43:1947
35. R Development Core Team (2004). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>
36. Fortran original by Leo Breiman, Adele Cutler, R port by Andy Liaw and Matthew Wiener. (2004). <http://www.stat.berkeley.edu/users/breiman/>
37. Clayden J, Greeves N, Warren S, Wothers P (2001) Organic Chemistry. Oxford University Press, Oxford