The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/Biodegradability Relationships

B. Cuissart,*[,†,‡] F. Touffet,[†] B. Crémilleux,[‡] R. Bureau,[†] and S. Rault[†]

Centre d'Études et de Recherche sur le Médicament de Normandie, UPRES EA 2126 5, rue Vaubénard, Université de Caen France, and Groupe de Recherche en Électronique Informatique et Imagerie de Caen, CNRS-UMR6072 Université de Caen, France

Received April 9, 2002

The maximum common structure between two molecules (MCS) induces a similarity that enables one to group compounds sharing the same pattern. This text relates a study based on such a structural depiction in a context of quantitative structure/biodegradability relationships (QSBR). The similarity indices are based exclusively on the MCS. First, the results of statistical tests prove that these indices significantly group compounds of similar activity together. These first conclusions enable the elaboration of classification models using those structural similarities. In a second part, a population of classifiers relying on the maximum common structure and the *k*-nearest-neighbor algorithm is explored. Finally, a thorough examination of the best models is conducted.

INTRODUCTION

The only way to reach a high level of protection for human health and environment from adverse effects of chemicals is an effective chemical policy. Recognizing that the current European Union (EU) policy does not provide sufficient protection and that a review was necessary, the European Commission made a commitment to assess the operation of four important legal instruments[1−4] governing chemicals in the community. The outcome was the adoption by the European Commission on February 13th, 2001, of a White Paper setting out the strategy for a future Community Policy for Chemicals.[5] Among priorities identified, it is mentioned that particular efforts need to be made for developing and validation of *in-vivo* and *in-vitro* test methods as well as modeling (QSAR, *Quantitative Structure Activity Relationships*). Indeed, an increasing place will be made in the future "REACH" (*Registration Evaluation and Authorization of Chemicals*) system for data predicted by models, especially for substances produced or imported in quantities between 1 and 10 tonnes per year. There is real challenge for researchers to develop modeling tools for the industry usable in a regulatory context. This is our goal by trying to develop new models for predicting biodegradability of chemicals.

A great amount of work has already focused on the field of Quantitative Structure-Biodegradability Relationships (QSBR), particularly by trying to validate existing models for legislative purposes[6−9] and then by setting several proposals to improve these models. The best models identified until now[10−13] and tested by Rorije et al.[9] are based on group contribution methods. Among these models, the one that could be easily acquired is BIODEG.[11] Its predictive power was evaluated on our data set and appears to give disappointed and insufficient results, not exposed here but available from the authors. These weaknessess justify the present work on QSBR. One of our research areas is the search of new descriptors. In this paper, a new kind of descriptor is presented and its interest in this field evaluated. First, the concept of maximum common structure (MCS) is described. Then the question of the relevance of a similarity measure based on MCS in the context of biodegradability prediction is raised. This study is carried out by a χ^2 goodness-of-fit test whose results enable the advancement of the study. Finally, a first classifier based on MCS similarity measure is built and the results are discussed.

MATERIAL AND METHODS

Biodegradation Data Sets. Two data sets, "MATE" and "BIOWIN", have been collected. The remaining "BOTH" is the merger of the two others. A summary of the three sets is displayed in Table 1.

The set MATE contains data from notification files of the French Ministry of Environment,[14] registered between 1994 and 2001 in the European Union. All the chemicals composed of a single identified chemical substance and whose readily biodegradability data are known have been introduced in the learning set MATE. A binary notation was adopted to distinguish readily biodegradable compounds (notation 1) from not biodegradable ones (notation 0) according to the regulatory criteria of the environmental hazard classification contained in the guide for the classification and labeling.[15,16] This categorization leads to 92 readily biodegradable compounds versus 899 which are not. Because of the confidentiality agreement signed between our laboratory and the Ministry, chemical structures cannot be displayed. An essential point is the data quality, which is essential to derive QSBR models. It has to be pointed out that the level of reliability of the biodegradation data of our data set is very high since all the data come from biodegradability tests

* Corresponding author phone: +33 (0)2 31 56 59 10; e-mail: cuissart@pharmacie.unicaen.fr.
† Centre d'Études et de Recherche sur le Médicament de Normandie.
‡ Groupe de Recherche en Électronique Informatique et Imagerie de Caen.

10.1021/ci020017w CCC: $22.00
Published on Web 09/23/2002

Table 1: Number of Compounds in the Learning Sets

set	MATE	BIOWIN	BOTH
biodegradable compounds	92	186	278
persistent compounds	899	109	1008
total	991	295	1286

carried out according to internationally validated test methods[17,18] and good laboratory practices.[19] Moreover, these data have been checked by competent authorities before to be included in European notification abstracts from which the data used in this study have been extracted.

The second data set, BIOWIN in Table 1, has been collected by P. H. Howard et al., with the intention of facilitating the development of structure/biodegradability relationships. This set, containing validated biodegradation data, is now a part of the Environmental Fate Data Base, available from Syracuse Research Corp. It has been used as the training set of the commercially available classification models, BIODEG,[11] implemented in the BIOWIN software. Data contained in BIOWIN have been evaluated by the authors[20] of this set and are acknowledged to be of high quality. The set BIOWIN is more balanced than MATE since 186 molecules are readily biodegradable against 109 which are not.

The Maximal Common Substructure. The presence of a common structural pattern in a set of molecules could be responsible for their similar behavior.[21] For this reason, the Maximal Common Substructure (MCS) has been used in structural depiction of molecules. For 25 years, this kind of depiction has entered in a wide variety of applications, from database screening[22] to unsupervised clustering with the CHARISMA package marketed by Tripos Inc.[23] There are several ways to define a common structure: connected,[24,21] disconnected,[26] partial,[27,21] or induced.[28] Here, we consider as a Maximal Common Substructure (MCS) between two structures, a connected induced part common to the two structures that contains the largest number of non-hydrogen atoms. This notion is exactly the same as in the work of Bayada et al.[24] It respects all the interatomic bonds, and it defines a common pattern between two molecules where chemical knowledge does not play any part.

The Calculation of the Maximal Common Substructure. The 2D structures are given through SMILES notation. To work directly with molecular graphs, an attributed grammar has been implemented to depict SMILES codes. Due to advances in hardware and algorithms, the calculation of a MCS between two molecular structures is now possible in a reasonable time.[25] We have implemented our own method of calculation (available in a technical report and subject of a forthcoming paper). The resulting program provides an exact solution quickly enough for the context of the experiment.

The Resulting Similarity Indices. Two indices have been defined to quantify the underlying similarity to further base our QSBR study on the MCS between two molecules. A molecule which biodegradability datum is used to compute the prediction model is named "Instance Structure" (IS). A molecule for which biodegradability datum is predicted by this model is named a "Query Structure" (QS). As we simply want to quantify the part that is shared between the two structures, the two similarity indices are defined as follows:[22]

$$\mathrm{Sim}_1(\mathrm{QS}, \mathrm{IS}) = \frac{|\mathrm{MCS}|}{|\mathrm{QS}|}$$

$$\mathrm{Sim}_2(\mathrm{QS}, \mathrm{IS}) = \frac{|\mathrm{MCS}|}{|\mathrm{QS}|} \times \frac{|\mathrm{MCS}|}{|\mathrm{IS}|}$$

with $|S|$ = the number of non-hydrogen atoms within S.

The index Sim_2 attaches the same importance to the two structures, whereas the index Sim_1 favors the Query Structure. For example, Figure 1 displays the structures of three acids. They all share the same pattern. Let us suppose that the terephthalic acid is the Query Structure. With the similarity index Sim_1, this molecule has the same similarity with the two other compounds, whereas with the index Sim_2, the terephthalic acid is closer to the aminobenzoic acid than to the nitrobenzoic acid.

The χ^2 Goodness-of-Fit Test and the Discrimating Power of the MCS. The ability of a similarity index to group compounds of similar activity may be called its discriminating power. Thanks to the similarity indices defined above, a set of compounds can be ranked according to the decreasing similarity toward a query structure. As the activity of the query structure is known, each compound within the set of instances may be considered as a success or a failure (see Figure 2). This process is repeated for a whole set of query structures. The related random phenomenom is ruled by an hypergeometric law with proper parameters. A way to assess if a similarity index groups compounds of similar activity is to compare the successes obtained within the top of these lists of neighbors with the expectation of the corresponding geometric law. We study the random variable X that corresponds to the sum of the successes within the first k compounds of a list. Thanks to the χ^2 goodness-of-fit test,[29] we are able to know whether X follows the corresponding random law or not. If X is proved to be independent of the corresponding hypergeometric law, the similarity index possesses a discriminating power.

To carry the χ^2 goodness-of-fit test out, the *empirical repartition* of the random variable X has to be calculated. When the first k elements of the lists of neighbors are considered, the empirical repartition of X can be seen as a table of $k + 1$ cells (from 0 to k). The ith cell indicates the number of lists within the sample that have i successes among the k first elements. The difficulty in the calculation of the empirical repartition of X lies in the fact that the elements within a list are not strictly ordered. A list can be considered as a succession of levels that may contain several compounds, each being equally similar to the query compound. Given a list L, when the sum of its successes X_L within the first k elements is counted, two cases may be encountered (see Figure 3). When k suits the end of a level, the successes within the first k elements are simply summed, and the proper counter of X is incremented accordingly. When k does not correspond to the end of a level (see Figure 4), it is necessary to carry out an additional calculation. Let p the number of elements to take into account within the jth level. As each element of this jth level has the same probability to be within the first p compounds, every combination has to be equally considered within this level. The selection of the first p

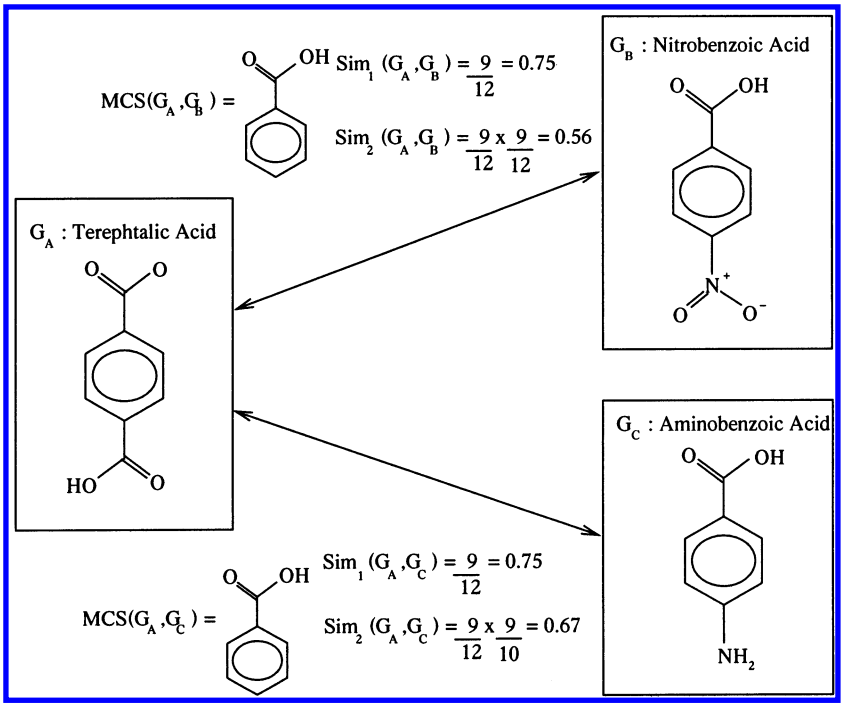

Figure 1. Examples of MCS and associated similarity indices.

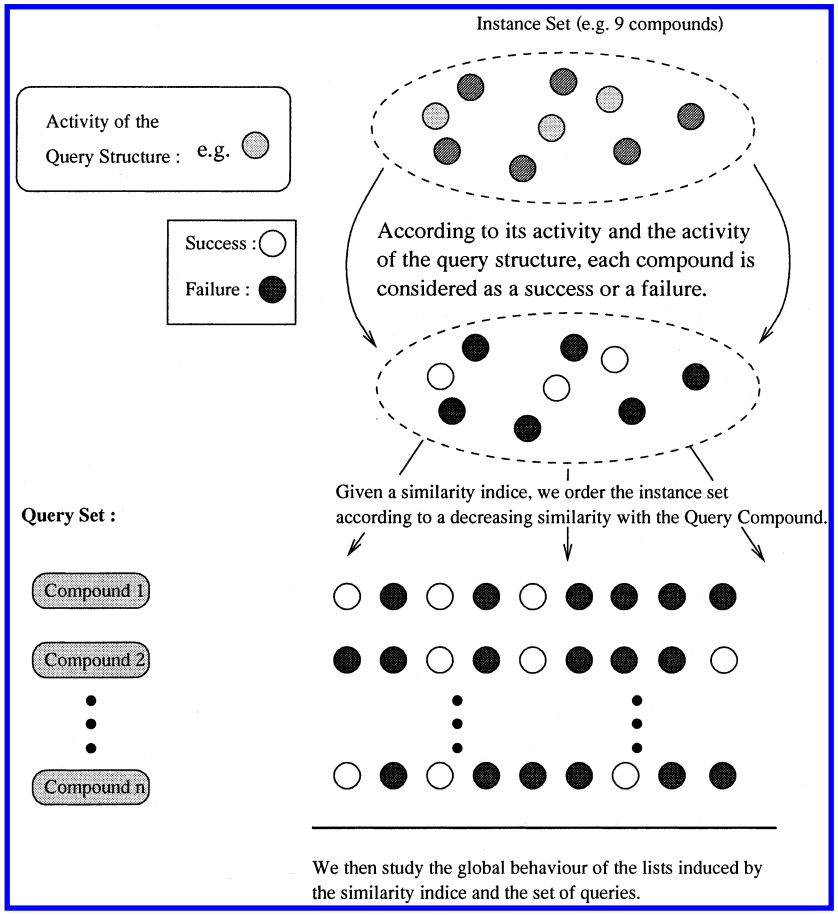

Figure 2. Ordering a set of instance compounds.

elements is ruled by an hypergeometric law. To know which values to add to the counters, we first sum the successes within the levels before the last level to take in account. The hypergeometric law enables the calculation of the probability to have *i* successes in the first *p* compounds of the *j*th level. These results indicate the proper values to add to the counters. This statistical tool assesses whether the similarities based on MCS group compounds of similar activity or not.

The *k*-Nearest Neighbor Classifiers. If the χ^2 goodness-of-fit tests show that the similarity indices Sim_1 and Sim_2 group compounds of similar biodegradability, it would be interesting to quantify this discriminating power. With this

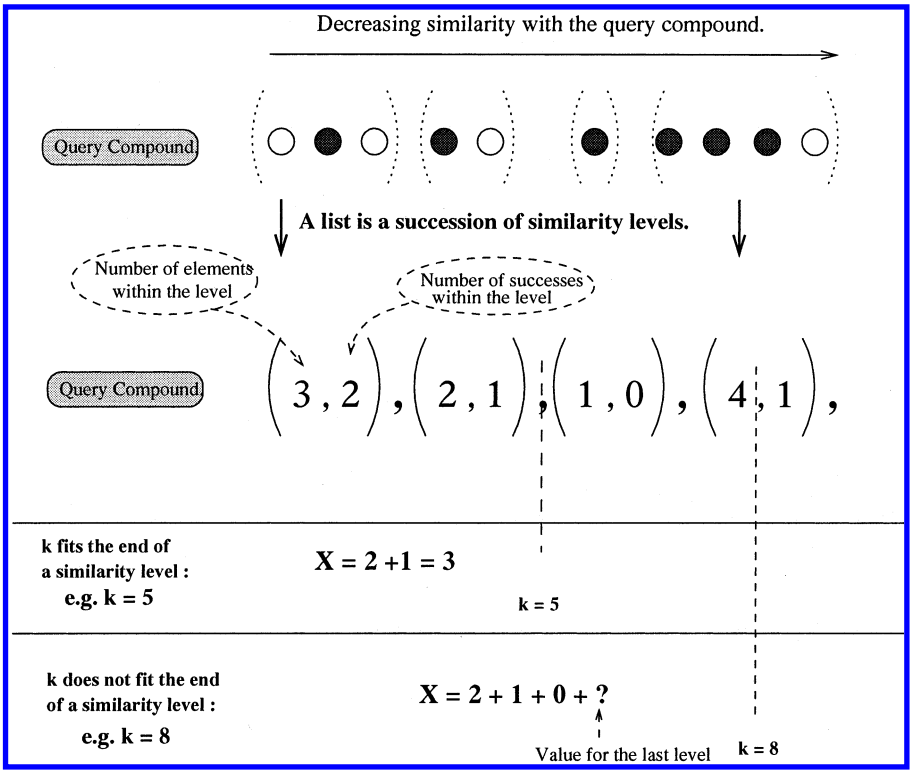

Figure 3. Counting of the successes within the *k* first elements of a list.

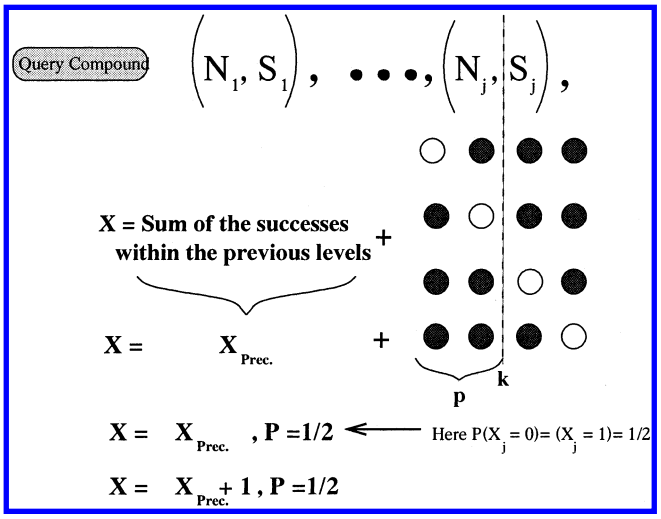

Figure 4. Counting the successes within a part of a level.

viewpoint, classifiers based only on MCS have been implemented and their performances have been evaluated. As the similarity indices allow the constitution of a neighborhood for a given Query Structure, it seems natural to build instance-based models of classification, relying on the *k*-nearest neighbor algorithm.[30]

The studied classifiers are based on rules described as follows. Given a Query Structure (QS) we aim to predict the activity, the model looks within a set of instance compounds and selects those whose similarity with QS exceeds a given threshold (*s*). A score function is then calculated by evaluating the ratio of biodegradable molecules to persistent ones within this selected neighborhood. If the score exceeds a given threshold (*d*), the Query Structure will be classified as biodegradable. Otherwise it will be estimated as not-biodegradable. Within this context, the search for the best model can be expressed as follows: given a set of query structures, a set of instances, find the pair (s_{max}, d_{max}) that maximizes the performances of all the considered rules.

The performances of a classification model are strongly dependent on the sets used as set of queries and as set of instances. Our learning set contains too few biodegradable compounds to set aside a validation set. Instead of discarding meaningful data, we prefer to assess the model by cross-validation. Leave-one-out cross-validation was chosen for two reasons.[31] First, the leave-one-out cross-validation estimates of prediction performances are approximately unbiaised from the true prediction ones.[31] Second, a leave-one-out approach is easily implemented in *k*-nearest-neighbor algorithm since no additional training effort is required each time the training set is redefined, avoiding a considerable computational burden.[32] Three learning sets were considered (MATE, BIOWIN, and BOTH) from which a query compound was repeatedly removed.

The next topic is the measurement of the performances of a given classification model. Given a model *M*, let

TP: the rate of biodegradable molecules well classified by the model *M*,

TN: the rate of not-biodegradable molecules well classified by the model *M*,

FN: the rate of biodegradable molecules that are classified as not-biodegradable by *M* ($TP + FN = 1$),

FP: the rate of not-biodegradable molecules that are classified as biodegradable ($TN + FP = 1$).

We choose to assess the quality of a model with a function $Q = \alpha \times TP + \beta \times TN$, where α and β are two coefficients that weight the importance to the classification of either the biodegradable compounds or the not-biodegradable ones. With this kind of quality function, the accuracy of a classifier can be visualized within the Receiver Operating Characteristic (ROC) space, a classic method from the signal detection theory that has been used in Artificial Intelligence classifier

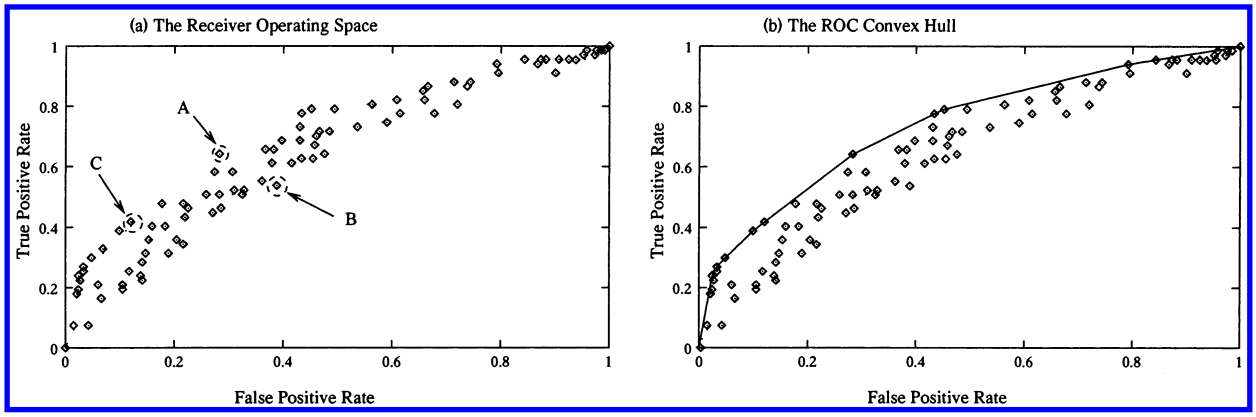

Figure 5. The receiver operating characteristic space and the ROC convex hull.

works.[33,34] In ROC space, a model M that has the performances (TP_M, TN_M, FN_M, FP_M) is projected on the point (FP_M, TP_M). The accuracy of a model can be assessed with its projection onto the ROC space: nearest the northwest corner (0,1) is its projection onto the ROC space, better is the model. Figure 5(a) displays the projections of the performances of numerous models on the ROC space (a diamond equals a projection). For example, the model B is obviously less accurate than A and C, whereas the choice between A and C proves to be more delicate.

As we cannot quantify the difference between the importance of a biodegradable compound and a not-biodegradable one, it is difficult to give a value to the parameters of Q, α, and β. As the function Q is not completely specified, the comparison of the models of classification has to be done under an *imprecise cost distribution*.[35] In this context, the notion of ROC convex hull (ROCCH) plays an important part.[36] Given a set of classification models, the ROCCH enables the selection of the subset of the best models whatever the weighting schemes (α,β) used in the quality function Q are. For example, the bold curve on the northwest of Figure 5(b) represents the ROC convex hull calculated from the set of points of Figure 5(a): it includes all the best classifiers. The main characteristics of the ROC convex hull are based on the two following properties. On one hand, if the projection of a classification model M belongs to the ROC convex hull, then the performances of M, (TP, TN, FN, FP) maximize the function Q for a weighting scheme (α_M, β_M). On the other hand, if the projection of M does not belong to the ROC convex hull, then the performances of M never maximize the function Q whatever the weighting scheme used. The ROC convex hull is a useful tool to exhibit the best models within a set of classifiers.

Given a set of query compounds and a set of instance compounds, the search of the best classifiers is conducted as follows. In a first step, for each pair of value (s, d), the resulting rates TP, TN, FP, and FN are measured. The program ROCCH[37] then calculates the corresponding ROC convex hull. Finally, the most interesting classifiers are selected among those lying on the hull.

RESULTS AND DISCUSSION

The χ^2 Goodness-of-Fit Test. The experiments related here aim to know whether the similarity indices Sim_1 and Sim_2 are able to group compounds of similar activity or not. There are six query sets, homogeneous according to the biodegradability property of the compounds (the biodegradable compounds respectively from BIOWIN, MATE, and BOTH and the not-biodegradable ones), three sets of instance compounds (BIOWIN, MATE, and BOTH), and two similarities. For each of these 36 cases, the mean of the sum of the successes obtained from the first k elements of the lists is compared with the expectancy resulting from the corresponding hypergeometric law. The diagram displayed in Figure 6 visualizes the ratio between the mean of the cumulated successes measured within our calculated lists and the expectancy of the corresponding hypergeometric law. On this example, the query set is constituted by the biodegradable molecules from BIOWIN, the instance set by all the compounds (BOTH from which we repeatedly remove the query compound) and the similarity index is Sim_2. For example, if we only consider the five nearest neighbors, the lists contain 3.5 times more biodegradable compounds than randomly constituted ones. This promotes this idea that the grouping is effective according to the activity. We observe the same type of representation for the other 35 calculations. So, it is graphically noticed that Sim_1 and Sim_2 seem to group compounds of similar activity. The χ^2 goodness-of-fit test is essential to assert the discriminating power of the similarity indices.

For each of the two indices, each of the six query sets, each of the three instance sets, a statistical test has been made to assert whether the similarity indices Sim_1 and Sim_2 are able to group compounds of similar activity or not. The lists are considered up to the 50th neighbor. The null hypothesis is H_0, "the lists could be ruled by the hypergeometric law", against the alternative H_1, "the lists are free from the hypergeometric law". The results are shown in Table 2. The displayed numbers represent the probability to correctly reject the null hypothesis. If the lists with more than the four nearest compounds are examined, all the 36 contexts reject the null hypothesis (see the repeated results of 99, 9%): these lists are not ruled by the corresponding hypergeometric law. The few results lower than 99.9% do not represent a relation between the lists and the random law. The results of these tests clearly prove that the indices group compounds of similar activity. Given the discriminating power of these similarity indices, the question of the efficiency of such similarities in a supervised classification context has to arise. At this point of the experimental design, no argument can be given to conclude whether one of the two indices, Sim_1 and Sim_2, is more efficient or not.

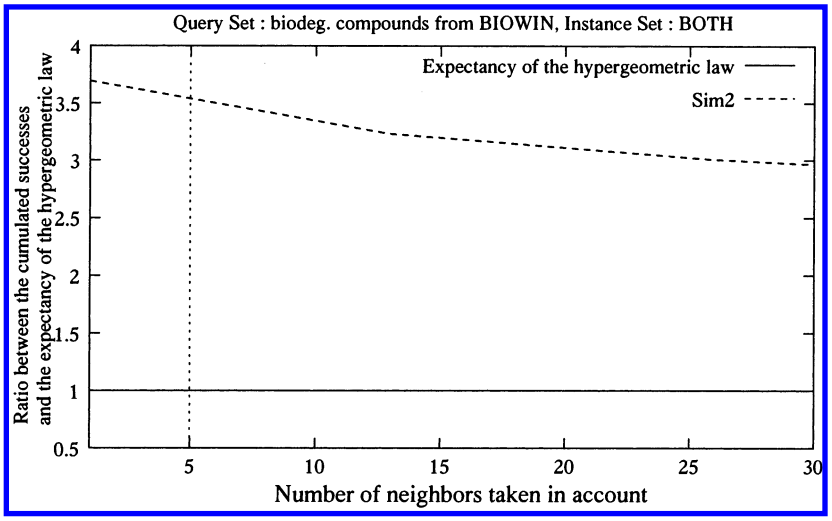

Figure 6. Graphical representation of the discriminating power.

Table 2. Results of the χ^2 Goodness-of-Fit Test

	activity	Query Set (QS)	|QS|	Instance Set (IS)	|IS|	similarity index Sim_1, %					similarity index Sim_2, %				
						1[a]	2	3	4	5−50	1	2	3	4	5−50
1	biodeg.	BIOWIN	186	BIOWIN	294(= 295 − 1)	99.9[b]	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
2	biodeg.	BIOWIN	186	MATE	991	<80	<80	80	99.9	99.9	99.9	99.9	99.9	99.9	99.9
3	biodeg.	BIOWIN	186	BOTH	1285(= 1286 − 1)	<80	80	97.5	99.9	99.9	99.9	99.9	99.9	99.9	99.9
4	biodeg.	MATE	92	BIOWIN	295	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
5	biodeg.	MATE	92	MATE	990(= 991 − 1)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
6	biodeg.	MATE	92	BOTH	1285(= 1286 − 1)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
7	biodeg.	BOTH	278	BIOWIN	294 *or* 295	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
8	biodeg.	BOTH	278	MATE	990 *or* 991	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
9	biodeg.	BOTH	278	BOTH	1285	95	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
10	not biodeg.	BIOWIN	109	BIOWIN	294	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
11	not biodeg.	BIOWIN	109	MATE	991	<80	90	95	97.5	99.9	<80	<80	99.9	99.9	99.9
12	not biodeg.	BIOWIN	109	BOTH	1285	99.9	99.9	99.9	99.9	99.9	<80	99.9	99.9	99.9	99.9
13	not biodeg.	MATE	899	BIOWIN	295	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
14	not biodeg.	MATE	899	MATE	990	99.9	99.9	99.9	99.9	99.9	80	99.9	99.9	99.9	99.9
15	not biodeg.	MATE	899	BOTH	1285	99.9	99.9	99.9	99.9	99.9	80	99.9	99.9	99.9	99.9
16	not biodeg.	BOTH	1008	BIOWIN	294 *or* 295	99.9	99.9	99.9	99.9	99.9	80	99.9	99.9	99.9	99.9
17	not biodeg.	BOTH	1008	MATE	990 *or* 991	99.9	99.9	99.9	99.9	99.9	80	99.9	99.9	99.9	99.9
18	not biodeg.	BOTH	1008	BOTH	1285	99.9	99.9	99.9	99.9	99.9	90	99.9	99.9	99.9	99.9

[a] Numbers of neighbors taken in account. [b] Probability that the null hypothese is false.

Table 3: Characteristics of the Models Belonging to the Convex Hull of Figure 7

identifier of the rule (indice-set- *n*°)	TN	TP	*s*	*d*
I-BI-1	1.00	0.00		
I-BI-2	0.91	0.45	0.60	0.80
I-BI-3	0.91	0.45	0.75	0.80
I-BI-4	0.89	0.56	0.55	0.70
I-BI-5	0.85	0.64	0.55	0.60
I-BI-6	0.79	0.82	0.75	0.40
I-BI-7	0.57	0.89	0.65	0.40
I-BI-8	0.4	0.95	0.55	0.35
I-BI-9	0.23	0.97	0.45	0.35
I-BI-10	0.13	0.99	0.35	0.40
I-BI-11	0.00	1.00		

First Experiments with the *k*-Nearest-Neighbor Method. Here are displayed the results related to the classification models based on the *k*-nearest-neighbor method and on the similarity indices Sim_1 and Sim_2. With all the possible combinations, there are 18 contexts of classification (two indices × three sets of query compounds × three sets of instances). As the models are evaluated with a leave-one-out validation, the query set has to be the same as the learning set. There are then only six classification contexts left (two indices × three sets of compounds). All the models were computed as follows. Given a query compound, the score function calculates the rate of biodegradable instances within the *decision support* (the molecules from the set of instances which similarity exceeds *s* with the query compound). If the rate of biodegradable compounds within the support is over *d*, the query compound will be predicted as a biodegradable compound and as a not-biodegradable in the opposite case.

Within a given context (a learning set and a similarity index), the best models of classification are selected thanks to the ROC convex hull. Figure 7 displays the convex hull related to the similarity index Sim_1 and the learning set BIOWIN. A first comment can be made by comparing the curve with the line which equation is "$y = x$". The selected models are effective if their convex hull is stretched toward the point (0,1). To go further in the examination of the results, the numerical results have to be directly analyzed. Table 3 lists all the best models, their rates TN and TP, and their parameters *s* and *d*. The convex hull always comprises two artificial models (I-BI-1 and I-BI-11 on the example) which represent respectively the following strategies: "all the

Table 4: Characteristics of All the Models Belonging to a ROC Convex Hull

similarity index Sim_1					similarity index Sim_2				
identifier	TN	TP	s	d	identifier	TN	TP	s	d
				Learning Set: MATE					
I-MA-1	1	0.0000	AllNeg		II-MA-1	1	0.0000	AllNeg	
I-MA-2	0.9933	0.0761	0.8	0.7	II-MA-2	0.9922	0.1304	0.8	0.55
I-MA-3	0.9933	0.0761	0.8	0.75	II-MA-3	0.9922	0.1304	0.8	0.6
I-MA-4	0.9933	0.0761	0.8	0.8	II-MA-4	0.9922	0.1304	0.8	0.65
I-MA-5	0.9844	0.1413	0.8	0.35	II-MA-5	0.9611	0.2391	0.65	0.35
I-MA-6	0.9644	0.2391	0.8	0.2	II-MA-6	0.9611	0.2391	0.65	0.4
I-MA-7	0.9466	0.3152	0.8	0.15	II-MA-7	0.9477	0.2826	0.65	0.1
I-MA-8	0.9288	0.3804	0.75	0.15	II-MA-8	0.9477	0.2826	0.65	0.15
I-MA-9	0.8376	0.5435	0.5	0.1	II-MA-9	0.9099	0.3696	0.45	0.3
I-MA-10	0.8131	0.5761	0.4	0.1	II-MA-10	0.802	0.5543	0.3	0.2
I-MA-11	0.7998	0.5870	0.3	0.1	II-MA-11	0.6085	0.7935	0.3	0.1
I-MA-12	0	1.0000	AllPos		II-MA-12	0	1.0000	AllPos	
				Learning Set: BIOWIN					
I-BI-1	1	0.0000	AllNeg		II-BI-1	1	0.0000	AllNeg	
I-BI-2	0.9174	0.4516	0.6	0.8	II-BI-2	0.9266	0.5484	0.45	0.8
I-BI-3	0.9174	0.4516	0.75	0.8	II-BI-3	0.9174	0.5860	0.5	0.8
I-BI-4	0.8899	0.5645	0.55	0.7	II-BI-4	0.8899	0.6667	0.75	0.75
I-BI-5	0.8532	0.6452	0.55	0.6	II-BI-5	0.8532	0.7419	0.75	0.55
I-BI-6	0.6881	0.8226	0.75	0.4	II-BI-6	0.7523	0.8441	0.5	0.65
I-BI-7	0.5688	0.8978	0.65	0.4	II-BI-7	0.5688	0.9247	0.4	0.55
I-BI-8	0.3945	0.9570	0.55	0.35	II-BI-8	0.4954	0.9516	0.4	0.5
I-BI-9	0.2202	0.9785	0.45	0.35	II-BI-9	0.3853	0.9677	0.35	0.5
I-BI-10	0.1284	0.9892	0.35	0.4	II-BI-10	0.1468	0.9892	0.3	0.35
I-BI-11	0	1.0000	AllPos		II-BI-11	0	1.0000	AllPos	
				Learning Set: BOTH					
I-BO-1	1	0.0000	AllNeg		II-BO-1	1	0.0000	AllNeg	
I-BO-2	0.9692	0.1871	0.8	0.3	II-BO-2	0.9633	0.4640	0.8	0.55
I-BO-3	0.9544	0.2770	0.8	0.25	II-BO-3	0.9633	0.4640	0.8	0.6
I-BO-4	0.8998	0.5288	0.8	0.15	II-BO-4	0.9524	0.5504	0.75	0.55
I-BO-5	0.8671	0.6439	0.8	0.1	II-BO-5	0.9315	0.5971	0.7	0.55
I-BO-6	0.8353	0.7122	0.75	0.1	II-BO-6	0.876	0.6763	0.7	0.2
I-BO-7	0.7024	0.8273	0.55	0.1	II-BO-7	0.8472	0.7086	0.65	0.3
I-BO-8	0.6171	0.8921	0.5	0.1	II-BO-8	0.7232	0.8165	0.5	0.35
I-BO-9	0.5704	0.9173	0.45	0.1	II-BO-9	0.5754	0.8921	0.4	0.3
I-BO-10	0.4861	0.9460	0.4	0.1	II-BO-10	0.4762	0.9317	0.4	0.15
I-BO-11	0.4167	0.9640	0.35	0.1	II-BO-11	0.4038	0.9568	0.35	0.15
I-BO-12	0	1.0000	AllPos		II-BO-12	0.3383	0.9676	0.3	0.2
					II-BO-13	0.3056	0.9712	0.3	0.15
					II-BO-14	0	1.0000	AllPos	

Table 5: Selected Models with BIOWIN or BOTH as Learning Set

learning set	similarity index	identifier	TP, %	TN, %	*s*	*d*
BIOWIN	Sim_1	I−BI-5	64.5	85.3	0.55	0.6
BIOWIN	Sim_2	II−BI-5	74.1	85.3	0.75	0.55
BOTH	Sim_1	I−BO-6	71.2	83.5	0.75	0.1
BOTH	Sim_2	II−B0−7	70.8	84.7	0.65	0.3

compounds are negative" and "all the compounds are positive". These two models are necessary for the design of the convex hull. There is no value for *s* and *d* associated with these artificial classifiers. For each context of study, the two models that minimize the difference between TP and TN are selected among those lying on the hull (I-BI-5 and I-BI-6 on the example). To go further in our selection of the best models, attention has been paid to limit the number of false positive as far as possible. Legislation promotes the development of readily biodegradable compounds since this kind of degradation data refers to complete mineralization of the original chemical. Indeed, keeping in mind that predictions of the models can be used in regulatory purposes it would always be better to have false negative instead of positive, for the protection of environment. For example, in the experiment, the model I-BI-6 is selected.

Table 6: Decision Supports of the Two Best Models with the Learning Set BOTH

	no. of compds	no. of empty supports	no. of nonempty supports	av size of the nonempty supports[a,b]	av rate of biodegradable compds[a]
			I-BO-6 (Sim_1)		
TP	198		198	247	0.199
FN	80	34	46	78	0.06
TN	842	396	446	28	0.049
FP	166		166	95	0.171
			II-BO-7 (Sim_2)		
TP	197		197	9.96	0.724
FN	81	59	22	3.77	0.168
TN	854	590	264	2.21	0.068
FP	154	-	154	5.83	0.628

[a] We only take in account nonempty supports. [b] We call average size of a support, the average number of instances above the similarity threshold *s*.

Table 4 displays all the models belonging to the ROC convex hull in one of the six studied contexts. A first examination of these results shows that the set MATE is more difficult to predict than BIOWIN or BOTH. On the first third of the table, there is only one relatively effective model (II-MA-11 reaches 60.8% of true negatives and 79.3%

Table 7: Prediction Performances of the Two Best Models with the Learning Set BOTH

	I-BO-6 (Sim_1), %			II-BO-7 (Sim_2), %		
queries from:	MATE	BIOWIN	BOTH	MATE	BIOWIN	BOTH
TP	55.43	79.03	71.22	40.21	86.02	70.86
FN	44.56	20.96	28.77	59.78	13.97	29.13
TN	86.09	62.38	83.53	87.98	57.79	84.72
FP	13.90	37.61	16.46	12.01	42.20	15.27

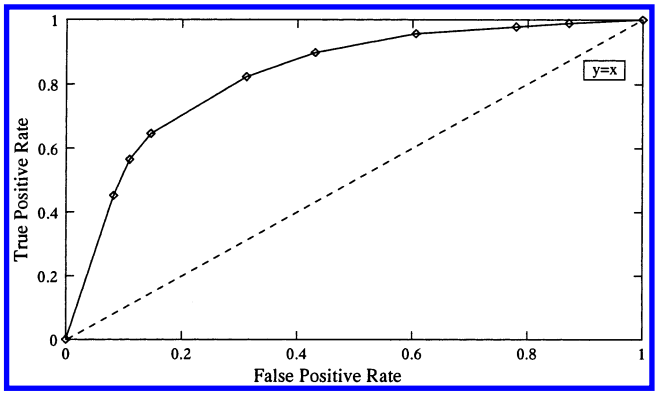

Figure 7. ROC convex hull collecting the best models for index Sim_1 and BIOWIN as learning set.

of true positives). These poor results may be explained by the repartition of the compounds within MATE: more than 90% of the compounds are not-biodegradable. There are not enough examples of biodegradable compounds to obtain a successful classifier just from MATE as set of instances. The set MATE has been excluded as a learning set. However, the information held in MATE is also contained in BOTH which is more well-balanced (21.6% of biodegradable compounds).

The two remaining learning sets are more well-balanced and their best models display interesting performances. Each of the four contexts brings one model out (see Table 5). These four models equally perform for the prediction of the not-biodegradable compounds as the classification rates concerning the biodegradable molecules show equivalent results. Keeping in mind that *s* is the threshold of minimum similarity of compounds included in the decision support, it is noteworthy that the *s* values associated with the best models are high. For example, the model I-BO-6 operates with $s = 0.75$: it means the decision is made with instances having more than 75% of their structure shared by the query compound. These models based on high-similarity comply with the principle of structure/activity relationship studies. The performances of the models are not different enough to argue that one of the two similarity indices is more efficient than the other one.

Thorough Examination of the Models Obtained with the Learning Set BOTH. The set BOTH, the merger of the two others, contains all the information on biodegradability and structures. The experiment described below intends to explain the classification mistakes by the examination of decision supports. The decision support of a Query Structure (QS) is the set of all the instance compounds which have a similarity exceeding *s* with QS. If the support is empty, the compound is predicted as not-biodegradable by default. If the support is not empty, the rate of biodegradable compounds is calculated within the support. To classify a compound as biodegradable, this rate has to be superior to the threshold *d*.

Details about decision supports are reported in Table 6. It can be seen that many query structures are predicted negative because their supports are empty (46% of the negatives for I-BO-6 and 69% of the negatives for II-BO-7). The average rate of biodegradable compounds within the nonempty supports is far higher for the positive cases than for the negative ones. With I-BO-6, the rates are respectively 0.199 and 0.171 against 0.06 and 0.049 (ratio of 1:3). With the model II-BO-7, this is more pronounced as the ratio reaches 1:4 (respectively 0.724 and 0.628 against 0.168 and 0.068). These comments support the discriminating power of similarity values based on MCS. Moreover, the rule I-BO-6, based on Sim_1, generates nonempty supports with very large average sizes (from 28 for the true negatives to 248 for the true positives). The model II-BO-7 produces supports with a smaller size (from 2 to 10 instances). With Sim_2, the decisions are based on the nearest neighbors, in compliance with the principle of the similar property.[38] Then, the discriminating power related to Sim_2 appears to be more important than the one related to Sim_1.

Table 7 distinguishes performances according to the origin set of the query (MATE, BIOWIN, or BOTH). The performances obtained with BOTH as query set are weighted means of the other ones. The main teaching of the table is that the biodegradable compounds from BIOWIN are better predicted than those from MATE (79% against 55% with I-BO-6 and 86% against 40% with II-BO-7). The phenomenom inverses for the not-biodegradable compounds: they are better classified when they come from MATE than from BIOWIN (86% against 62% with I-BO-6 and 88% against 58% with II-BO-7). An investigation has been conducted on these differences. The fluctuation of the performances may result from a factor of molecular diversity. As the indices Sim_1 and Sim_2 may group structures of similar molecular size, investigations were made to explore the size distribution.

Investigations on the Influence of the Molecular Size. The distribution of the sizes of the compounds is displayed in Figure 8. It can be noticed that the biodegradable compounds are smaller than the not-biodegradable ones, in conformity with the scientific knowledge.[11] Indeed, there are very few biodegradable structures with more than 30 non-hydrogen atoms large. Then, the problem may be addressed if there are enough instances of large biodegradable structures to correctly classify a large biodegradable query.

An experiment has been conducted with the index Sim_2 by analyzing the performances of the set of instances when the queries are restricted to compounds having more than 30 atoms (300 compounds, among which eight are biodegradable and 292 not-biodegradable). Figure 9 displays the convex hull related to this set of queries. The best model is isolated in the ROC space, and it performs only 62.5% percent of true positive cases. This experiment confirms the hypothesis that the information contained in the learning set is not sufficient to classify the compounds with more than 30 non-hydrogen atoms.

A Model of Classification That Takes into Account the Size of the Query Structure. The investigations have been followed by computing a classification model that takes the size parameter into account. Given the preceding remarks, Sim_2 has been used as similarity index, and the queries have

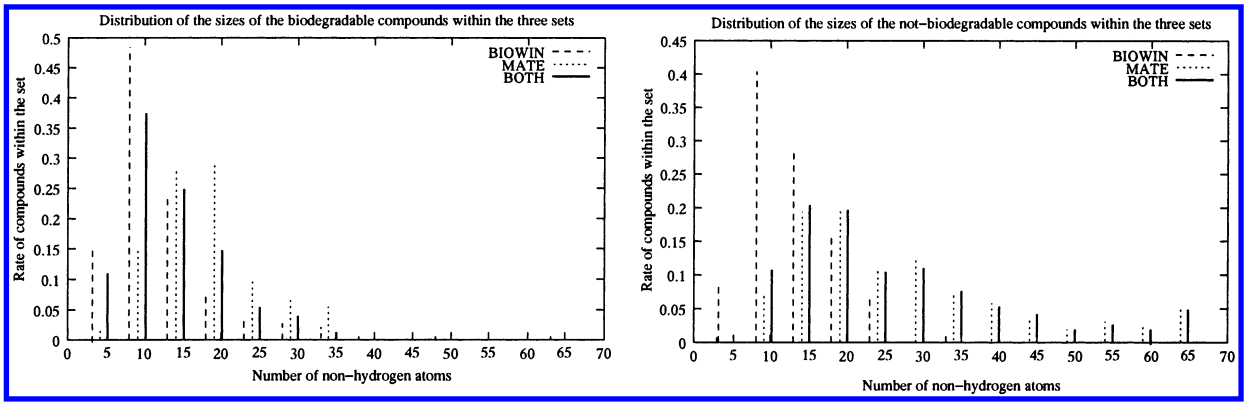

Figure 8. Distribution of the compounds according to their source, their activity, and their size.

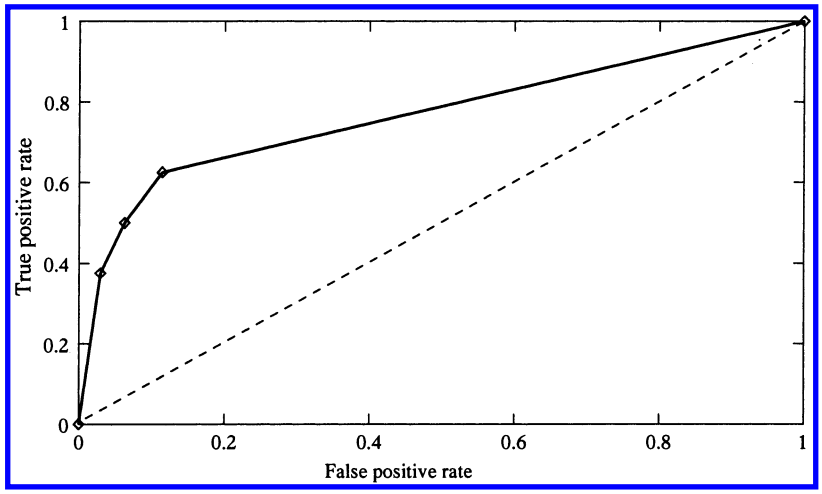

Figure 9. ROC Convex Hull related to the compounds with more than 30 non-hydrogen atoms.

been limited to the structures whose size is less than or equal to 30 non-hydrogen atoms large. The studied classifiers operate as follows. Given a query structure of size s_q, the subset of all the molecules which size is between $(s_q - 3)$ and $(s_q + 3)$ is extracted from the set of instances (BOTH from which we remove the query structure). This set is considered as a new query set (QS2). The parameters s_{max} and d_{max} which maximize the performances of classification for QS2 are then calculated. These parameters s_{max} and d_{max} are finally used to classify the query structure. The performances of this resulting model are $TP = 0.729$ and $TN = 0.819$. These results are similar to the model II-BO-7 ($TP = 0.708$ and $TN = 0.847$).

CONCLUSION AND PERSPECTIVES

To increase the quality of prediction tools in chemistry, the task consists of improving classification methods, in using them advisedly as well as in exhibiting the relevant information from the chemical structures. All this has to be done keeping in mind that a good tool not only works well for predictive purposes but provides a model that is understandable in terms of chemical knowledge. The maximum common substructure constitutes a natural descriptor for structural similarity between chemical compounds. It is well-known that the determination of the MCS between two structures requires a large amount of calculation. With the progress made on algorithms and hardware, this can be done in a reasonable time for medium sets of molecules. With the results of the χ^2 goodness-of-fit tests, it was demonstrated that the similarity indices Sim_1 and Sim_2 are discriminant to a group of compounds of similar activity. To assess whether the neighborhoods induced by the MCS are strong enough to be useful in a classification context or not, instance-based classifiers have been computed. The best models have been exhibited by the ROC convex hull and display very interesting results. The success rates prove that the MCS descriptor will be useful in a classification context.

The MCS enables a first investigation of a new context of classification. As no specific knowledge is required, a similarity calculation based on MCS enables an exploration of the training sets. For example, in our experiment, it was noticed that the set MATE contains too few biodegradable molecules to be a reasonable set of instances. The small number of large biodegradable structures does not allow the prediction of the activity of the large compounds. It can be concluded that MCS may be used as a basis for the elaboration of instance-based classifiers that directly operate on 2D structures.

Working on the way to quantify the similarity between the structures may significantly improve the performances of the classifiers. Without any specific knowledge, our experiment shows that the index Sim_2 operates more in conformity with the SAR principles than Sim_1. We intend to explore several ways in our future work. One way would be based on the notion of reduced graph where important structural features have been detected.[39,40] It may be possible to quantify the MCS in a more elaborate way than counting the non-hydrogen atoms. The efficiency of a similarity based

on MCS should increase with the intervention of knowledge related to the supervising activity. However, it seems more appropriate to associate a description based on MCS with other structural attributes. We plan to use a similarity based on MCS within an attribute-value depiction of the compound. For example, we select a set of structures, each typical of a modality of the supervising activity. In the process of classification, a high similarity with one of these structures influences the value of the target function. Then, the comparison between the performances of MCS-based similarity and those of a fragment-based one could be the subject of a study. A MCS-based similarity aims to improve other approaches rather than compete with them.

ACKNOWLEDGMENT

The authors thanks the Conseil Régional de Basse-Normandie, the company ATOFINA (Dr P. Lemaire) for their financial support, and the French ministry of environnment (Mrs. L. Musset) for its collaboration.

REFERENCES AND NOTES

(1) Council Directive 67/548/EEC relating to the classification, packaging and labeling of dangerous substances, as amended [OJ 196, 16.8.1967].
(2) Directive 88/379/EEC relating to the classification, packaging and labeling of dangerous preparations [OJ L 187, 16.7.1988].
(3) Council Regulation (EEC) 793/93 on evaluation and control of risks of existing substances [OJ L 84, 5.4.1993].
(4) Directive 76/769/EEC relating to restrictions on the marketing and use of certain dangerous substances and preparations [OJ L 262, 27.9.1976].
(5) White paper Strategy for a future Chemicals Policy [OJ 88, 27.2.2001].
(6) Structure−activity relationships for biodegradation. Environment monograph no. 68, Organisation for Economic Cooperation and Development, Paris, 1993.
(7) Loonen, H.; Lindgren, F.; Hansen, B.; Karcher, W. Prediction of biodegradability from chemical structure. In *Biodegradability prediction*; Peijnenburg, W. J. G. M., Damborsky J. I., Eds.; Kluwer Academic Publishers: 1996; pp 105−113.
(8) Langenberg, J. H.; Peijnenburg, W. J. G. M.; Rorije, E. On the usefulness and reliability of existing QSBRs for risk assessment and priority setting. *SAR QSAR Environ. Res.* **1996**, *5*, 1−16.
(9) Rorije, E.; Loonen, H.; Muller, M.; Klopman, G.; Peijnenburg, W. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere* **1999**, *38*, 1409−1417.
(10) Howard, P. H.; Boethling, R. S.; Stiteler, W. M. Predictive Model for Aerobic Biodegradability Developed from a File of Evaluated Biodegradation Data. *Environ. Toxicol. Chem.* **1992**, *11*, 593−603.
(11) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W. M.; Beauman, J.; Tirado, N. Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459−465.
(12) Loonen, H.; Lindgren, F.; Hansen, B.; Karcher, W.; Niemela, J.; Hiromatsu, K.; Takatsuki, M.; Peijnenburg, W.; Rorije, E.; Struijs, J. Prediction of biodegradability from chemical structure: Modeling of ready biodegradation test data. *Environ. Toxicol. Chem.* **1999**, *18*, 1763−1768.
(13) Klopman, G. The META − CASETOX system. In *Biodegradability Prediction*; Peijnenburg, W. J. G. M., Damborsky J. I., Eds.; Kluwer Academic Publishers: 1996; pp 27−40.
(14) Council Directive 92/32/EEC amending for the seventh time Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relative to the classification, packaging and labeling of dangerous substances, as amended [OJ L154, 5.6.1992].
(15) Commission Directive 93/21/EEC adapting to technical progress for the 18th time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances [OJ L 110, 4.5.1993].
(16) Commission Directive 96/54/EC adapting to technical progress for the twenty-second time Council Directive 67/548/EEC on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances [OJ L 248, 30.09.1996].
(17) Commission Directive 92/69/EEC to technical progress for the seventeenth time Council Directive 67/548/EEC on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances [OJ L 383, 29.12.].
(18) OECD Guidelines for the Testing of Chemicals, CD-ROM, OECD, Paris, 1998.
(19) The OECD Principles of Good Laboratory Practice, OECD, Paris, 1998.
(20) Howard, P. H.; Hueber, A. E.; Boethling, R. S. Biodegradation Data Evaluation for Structure/Biodegradability Relations. *Environ. Toxicol. Chem.* **1987**, *6*, 1−10.
(21) Takahashi, Y.; Satoh, Y.; Sasaki, S. Recognition of the Largest Common Fragment among a Variety of Chemical Structures. *Anal. Sci.* **1987**, *3*, 23−28.
(22) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.
(23) <http://www.tripos.com>.
(24) Bayada, D. M.; Simpson, R. W.; Johnson, A. P. An Algorithm for the Multiple Common Subgraph Problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680−685.
(25) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.
(26) McGregor, J. J.; Willett, P. Use of a maximal common subgraph algorithm in the automatic identification of the ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137−140.
(27) McGregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software-Practice Experience* **1982**, *42*, 23−34.
(28) Levi, G. A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo* **1972**, *9*, 341−352.
(29) Dacunha-Castelle, D.; Duflo, M. Problèmes à temps fixe. In *Probabilités et Statistiques*; Ciarlet, P. G., Lions, J. L., Eds.; Masson, 1994; pp 134−138.
(30) Dasarathy B. V. *Nearest Neighbor (NN) Norms: NN Patterns Classification Techniques*. IEEE Computer Society Press: 1991.
(31) Hastie, T.; Tobshirani, R.; Friedman J. Model Assessment and Selection. In *The Elements of Statistical Learning*; Springer-Verlag: 2001; pp 214−217.
(32) Moore, A. W.; Lee, M. S. Efficient Algorithms for Minimizing Cross-Validation Error. In *Proceedings of the eleventh International Conference on Machine Learning*; Morgan Kaufmann: 1994.
(33) Egan, J. P. *Signal Detection Theory and ROC Analysis*; Academic Press: 1975.
(34) Friedman, C.; Wyatt, J. *Evaluation methods in medical informatics*; Springer-Verlag: 1997.
(35) Provost, F.; Fawcett, T. Analysis and Vizualisation of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*; American Association for Artificial Intelligence: 1997.
(36) Provost, F.; Fawcett, T. Robust Classification for Imprecise Environments. *Machine Learning J.* **2001**, *42*.
(37) Fawcett T. <http://www.hpl.hp.com/personal/Tom_Fawcett/ROCCH/index.html>. GNU General Public License.
(38) Johnson, M.; Maggiora, G. *Concepts and applications of molecular similarity*; Wiley-Intersciences: 1990.
(39) Yuan, S.; Zheng, C.; Zhao, X.; Zeng F. Identification of maximal common substructures in structure/activity studies. *Anal. Chim. Acta* **1990**, *235*, 239−241.
(40) Takahashi, Y.; Sukekawa, M.; Sasaka, S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−643.

CI020017W