Relevance of Molecular Similarities based on the Maximum Common Substructure

Bertrand Cuissart[a,b], Frédérique Touffet[a], Bruno Crémilleux[b] and Jean-Jacques Hébrard[b]

[a] *Centre d'Études et de Recherche sur le Médicament de Normandie, UPRES EA 2126*
University of Caen
France
[b] *Groupe de Recherche en Électronique, Informatique et Imagerie de Caen, CNRS-UMR6072*
University of Caen,
France

Introduction

The expansion of the area of structure-activity relationships (SAR) is obviously rooted in the development of data processing techniques. In order to increase the quality of prediction tools in chemistry, the task consists in improving classification methods, in using them advisedly, as much as in exhibiting relevant information from chemical structures. This text intends to assess the efficiency of similarity measures based on the Maximum Common Substructure (MCS), appraising their ability to bring together compounds with similar behaviours. The notion of MCS will first be introduced. Then the interest of structural depictions based on MCS will be evaluated with an experiment on biodegradability data.

1 The Maximum Common Substructure : a natural but particular structural descriptor

Establishing SAR from 2D chemical structures does not only require the right method of analysis to extract knowledge, the most reliable biological data but also the exhibition of relevant features from the molecules. Parallel to advances in both computer hardware and computer sciences, the ways to depict a chemical structure multiply from the indices to relational representations (*e.g.* the feature trees [6]). Finding the Maximum Common Substructure between two molecular graphs lies in a class of problems that demand a large amount of calculations to be solved. Due to some particularities of molecular graphs, we are now able to find such an information quickly enough to exploit it.

Definition of the Maximum common substructure (MCS) The definitions are close to the notion used by D. M. Bayada and *al* in [1].

A molecular graph : we represent the atoms and bonds of a molecule as a labeled graph $G(V, E, L_V, L_E)$ where :

V : the set of vertices corresponding to the atoms,

E : the set of edges corresponding to the bonds,

L_V : a function that represents the atomic number,

L_E : a function that represents the bonding type.

A subgraph : we consider here a subgraph as a connected part of a graph that respects all the existing bonds.

A Maximum Common Substructure (MCS) ; Let's consider the size of a graph as the number of its vertices. We will say that a graph G_1 is a MCS between two graphs G_A and G_B if (see Fig. 1) :

i) G_1 is a subgraph for both G_A and G_B,

ii) there is no graph which is a subgraph for both G_A and G_B, including strictly more vertices than G_1.

Notes :

- Given two graphs, G_A and G_B, there is always a MCS between G_A and G_B (even the empty graph),
- the MCS between two graphs is not necessarily unique.

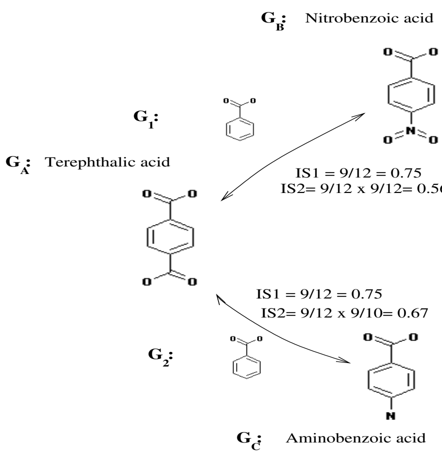
Quantifying the MCS with a similarity indice Afterwards, we will use MCS as a basis to assess structural similarity between two molecules. Precisely we will have a reference compound and we wish to order a set of test compounds according to their similarity with this reference molecule. In order to do that, we have to quantify the structural likeness figured by MCS. Under these designations, there are two obvious indices (see Fig. 1) :

$$I_{S_1} = \frac{|\text{MCS}|}{|\text{Reference}|},\ I_{S_2} = \frac{|\text{MCS}|^2}{|\text{Reference}| \times |\text{Test}|}\ (|.| = \text{number of atoms})$$

First notes on these similarities One can notice that MCS is somewhat a special basis for working a structural similarity out. Its main characteristics are :

- **it is a global similarity** : the calculus of the MCS does not attach any particular value to specific parts of the chemicals. Compounds are considered in their global nature,

Figure 1: The Maximum Common Substructure and the similarity indices



- **it is a relative similarity** : the indices defined above only indicate similarity between two compounds,
- **it does not require any expert knowledge** : when chemicals are represented with vectors, a feature selection is necessary. The MCS allows to work directly on structures, without *expert knowledge*. It's an important point that enables the use of MCS-based similarity in a wide range of experiences.

Other works based on MCS Although the MCS is not a common attribute used to describe chemical structures, there has been many reports of works based on this feature for 25 years. We choose to mention here three significant and recent examples of works using MCS as a basis for similarity.

The first experiment is provided by Hagadone's work on substructure similarity searching [3]. This work consists in extracting a selection of compounds from a chemical database. A fragment based similarity is used to calculate an upperbound to the size of the MCS between the target and each database structure. The molecules of the database are then ordered according to these upperbounds. Finally, a MCS is determined between the target and each compound of a first selection. The MCS search is based on a rapid but approximate subgraph algorithm. The author does not really assess the quality of his similarity, he measures the efficiency of his method through its ability to retrieve a target compound.

The second experiment mentioned is an example of SAR based on MCS [1]. The notion of MCS is exactly the same as our's. The authors explain a maximal common subgraph algorithm that enables the calculation of a MCS between two molecules, then between several compounds. Finally, they discuss about the performances of their algorithms with three examples, not on the quality of the similarities produced. The resulting program was significantly quicker than the previous published works on MCS.

The third example is a hierarchical clustering tool proposed by Tripos [5], CHARISMA. Clustering means classification not supervised by an activity. This program groups molecules according to their MCS. The result is a dendogram where each node represents a structure common to all compounds placed below. According to its authors, this package enables the obtention of interesting clusters, different from those obtained by others chemical clustering tools.

As we can see, the MCS has already been used in several works. But none of these experiments focused on the evaluation of the quality of MCS as a similarity feature. That is our goal in this text.

2 Experimental contribution for the evaluation of MCS-based similarities

The aim We wish to study the significance of the MCS as a part of the structural depiction of a compound. It is hard to appraise the interest of a structural description through a classification experiment. The analysis turns very complicated due to the interferences of the features and parameters of the selected method. In such a study, the first question should be : do the attributes group compounds of similar activity ? In order to answer this question, one has to free from any classification technique. That is what we intend to do with the experiment related here. In order to do that, we will describe a reference compound by ordering a set of molecules according to the size of their MCS with the reference compound. We will then compare this description with the one obtained by randomly ordering the same set of compounds. On the basis of this comparison, we will assess the ability of the MCS to group compounds of similar behaviours.

Material of the experiment For the experience, we have studied two sets of compounds, with both 2D structure (SMILES code) and ready biodegradability notified. The biodegradability is notified as a dichotomic activity ; its value is 1 if the corresponding compound is readily biodegradable, 0 otherwise. Here is a description of each set :

- 295 molecules used to work the BIOWIN model out [2]. Due do the precautions taken for the formation of the set [4], it can be considered as very reliable. We call this set **BIOWIN**,
- 344 molecules extracted from the "notification files" of the Ministère de l'Aménagement du Territoire et de l'Environnement, highly reliable data. We call this set **MATE**,
- we have considered a third set of 639 compounds, the gathering of BIOWIN and MATE. We call it **BOTH**.

Table 1 *Distribution inside the three datasets*

Set	*Biodegradable compounds*	*Not-biodeg. comp.*	*Total*
MATE	$54\ (15,7\%)$ *(set 1)*	$290\ (84,3\%)$ *(set 2)*	344
BIOWIN	$186\ (63\%)$ *(set3)*	$109\ (37\%)$ *(set 4)*	295
BOTH	$240(37,5\%)$	$399(62,5\%)$	639

The table 1 shows that the sets are very different. That is why we will carry on with considering BIOWIN and MATE as distinct sets.

In order to conduct the experiment, we have at our disposal some softwares, locally carried out :

- to work directly with molecular graphs, we have implemented an attributed grammar able to depict SMILES codes. We also have developed a tool that automatically codes a molecular structure as a SMILES code,

- to calculate an induced connected MCS between two molecular graphs, we have conceived an algorithm based on a specification of the search as a constraints satisfaction problem. The resulting program provides an exact solution quickly enough for the context of the experiment : a MCS between two compounds of medium size (number of atoms ≤ 100) is found within far less a second on a PC with a Pentium II.

The protocol We consider :

- four homogeneous sets of reference compounds : the sets 1, 2, 3, 4 of table 1,

- three test bases : BIOWIN, MATE and BOTH,

- two similarity indices I_{S_1} and I_{S_2}.

We repeat the following experiment for each group of reference molecules, each test base and each similarity indice (24 experiments) (see Fig 2):

i) for each molecule of the reference set (called the reference molecule), we order the test base according to the similarity (based on MCS) between each compound of the test base and the reference molecule,

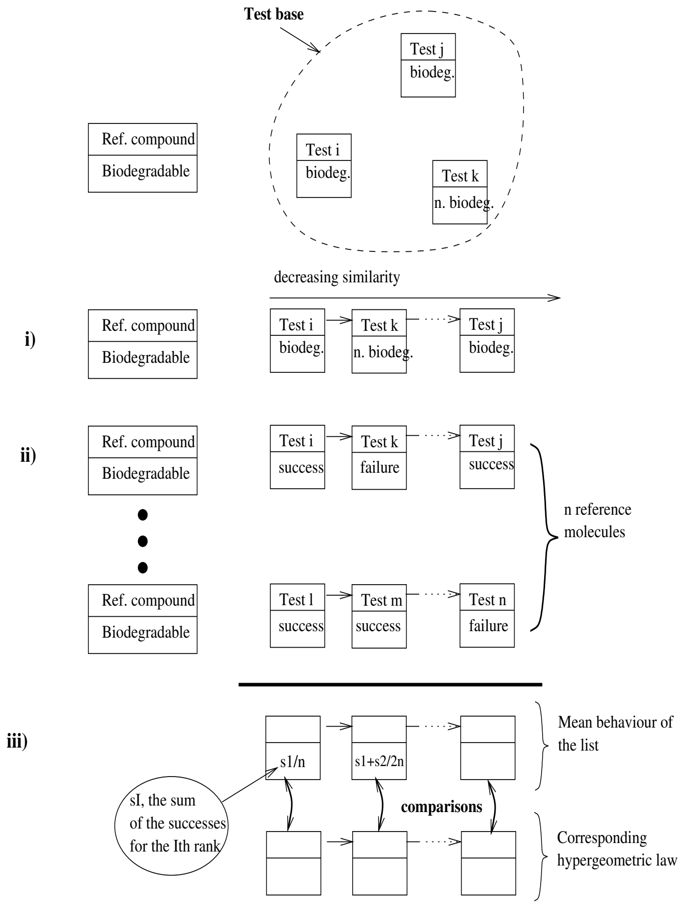
ii) within a reference group, all the compounds belong to the same class of biodegradability. We now consider each test compound as a success if it belongs to the same class of biodegradability as the reference molecule or as a failure otherwise. For each reference compound, we now have an ordered draw where each rank is labelled as a success or a failure,

iii) we compare the mean cumulated success obtained on the base of our similarity indice with the expectancy of success given by the hypergeometric law that rules the corresponding random draw. Here we have a measure to assess the ability of our similarity to group compounds with the same biodegradability. In order to assess the discrimant power of our indices, the significant measure can be expressed as follows :

$$Q_k = \frac{\text{Mean}_k}{\text{Expectancy}(k, n, p)} \text{ with :}$$

- Mean_k, the mean of the successes within a group of list,

Figure 2: Illustration of the main steps of the experiment



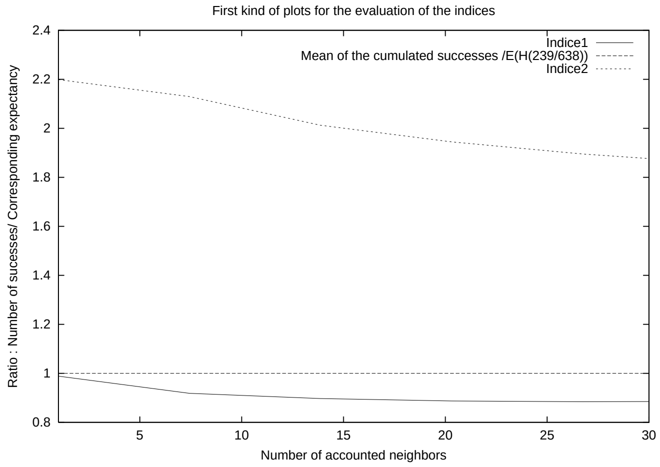
- $\text{Expectancy}(k, n, p) = k \times \frac{p}{n}$, the expectancy of the corresponding hypergeometric law, where n equals the number of ranked compounds, p the number of compounds considered like a success.

In a first time, our estimation of the quality of the indices will only depend on the value of this quotient. We also compare the discriminant powers of our two similarity indices. That is the reason why their quality measures will lie on the same plot. We have obtained 12 such plots.

3 Results and Comments

We have to keep in mind that we wonder if the compounds placed in the top ranks of a list possess an activity similar to the reference molecule. So we just plot the comparisons for the 30 first compounds, focalizing on the interesting phenomenom. We won't present here the 12 graphics. However, it would not have been very interesting. We decided just to show one example of the results we obtained. In this case, the reference compounds are the 186 biodegradable molecules from BIOWIN and the ordered base is BOTH, the reunion of the two others.

Figure 3: An illustration of the graphics



Within the 24 graphs (as for the graph of Fig. 3), there are 15 graphs that reflect a significant discriminant power of the indices. We also want to compare the two indices. At the reading of the table 2, the indice 2 apears to better discriminate when the reference is biodegradable. The first indice seems to give better results when the referent compound is not biodegradable. At this point of our experiment, we have calculated an average behaviour within lists determined in accordance with homogeneous referent compounds. To carry out our task, it appears very important to know how different are the lists within a group. In order to go further in the apprehension of our lists, we have calculated the standard deviation within each group of lists.

Table 2 (Comparisons of the lists calculated with the two indices) *The lists are only observed for the first neighbors ($k \leq 30$). A significant difference between the Q_k of the two lists is reported in the column of the concerned indice.*

Reference compound/Test base	*Indice* 1	*Indice* 2
Biodegradables from Biowin / Biowin		$+0,15$
Biodegradables from Biowin / MATE		$+0,2$
Biodegradables from Biowin / Deux		$> 1,0$
Not-biodeg. from Biowin / Biowin	$+0,3$	
Not-biodeg. from Biowin / MATE		
Not-biodeg. from Biowin / Deux	$+0,3$	
Biodegradables from MATE / Biowin		$+0,15$
Biodegradables from MATE / MATE		$> 1,0$
Biodegradables from MATE / Deux		$+1,0$
Not-biodeg. from MATE / Biowin	$+0,5$	
Non-biodeg. from MATE / MATE		
Non-biodeg. from MATE / Deux	$+0,5$	

4 Conclusion and Perspectives

We won't give here definitive conclusions on the importance of the MCS as a structural attribute. Our experiments point that :

- the similarity indices based on MCS have a significant discriminating power in the large majority of the cases,
- the two indices defined are not equally efficient. The second indice appears to be better to group biodegradable compounds and the other one more convenient for not biodegradable ones.

At the present time, these conclusions are not definitive and the challenge will be now to extend our experiments by :

- using continuous measures instead of a dichotomic one,
- proceeding statistical tests. For example, we would like to answer to questions like : are n lists sufficient to observe a discriminant power ?

However, our first conclusions are encouraging and we can envisage the use of the MCS as a base for structural description, either to screen a database, or to outline a new compound. For example,

- thanks to a nearest neighbors method, we could extract compounds, just like in the experiment of the text. According to their profiles, we could infer an outline for the unknown compound,
- we could also preselect representative structures and depict a new molecule as a vector where each attribute could be the similarity, based on MCS, between this target compound and the preselected ones.

The most promising way seems to consider the MCS as one kind of attribute inside an heterogeneous vector. This kind of representation will need efforts to be mastered. This is a probable way to enrich significantly the structural depiction of a compound and to increase the interest of the chemical databases.

References

[1] Denis M. Bayada, Richard W. Simpson, and A. Peter Johnson. An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.*, 32:680–685, 1992.

[2] R.S. Boethling, P.H. Howard, W. Meylan, W.Stiteler, J. Beauman, and N.Tirado. Group contribution method for predicting probability and rate of aerobic biodegradation. *Environ. Sci. Technol.*, 1994.

[3] T. R. Hagadone. Molecular substructure similarity searching : Efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.*, 32:515–521, 1992.

[4] P.H. Howard, A.E. Hueber, and R.S. Boethling. Biodegradation data evaluation for structure biodegradability relations. *Environmental toxicology and chemistry*, 1987.

[5] Tripos Inc. http://www.tripos.com.

[6] M. Rarey and J.S. Dixon. Feature trees : A new molecular similarity measure based on tree mathching. *Journal of Computer-Aided Molecular Design*, 12:471–490, 1998.

To contact us : cuissart@pharmacie.unicaen.fr,

Aknowledgements : The studies have been financed both by the Conseil Régional de Basse-Normandie and the company ATOFINA.