Pergamon

Tetrahedron, Vol. 53, No. 41, pp. 14085-14106, 1997
© 1997 Elsevier Science Ltd
All rights reserved. Printed in Great Britain
0040-4020/97 $17.00 + 0.00

PII: S0040-4020(97)00911-3

Classification of Organic Reactions Using Similarity

Guido Sello*, Manuela Termini

Dipartimento di Chimica Organica e Industriale, Universita' degli Studi di Milano, via Venezian 21, 20133 Milano, Italia

Abstract: Organic reaction classification allows for better understanding of organic reactivity, better data sharing, and better reaction storage and retrieval. The power of similarity has been often used to strengthen search methodology and to help reaction prediction systems, but seldom to optimise reaction classification. A hierarchical classification by similarity measures is proposed. It is based on electronic energy and chemical potential descriptors and on a general description of reactions. Using a main division into three reaction sets, additions, eliminations, and substitutions, followed by two successive subdivisions by number and types of reactive atoms it is possible to arrive at a hierarchy of groups containing homogeneous reaction classes. In addition, inside each set the reactions can be ordered. Some simple examples are used to demonstrate the efficiency of the classification and its overall consistency. The application of the final system to more complex examples is further proof of its efficiency. The reported classification can also be used in synthesis planning.

INTRODUCTION

Since many years the classification of organic reactions has represented a fascinating argument for organic chemists.[1] The appeal was the result of the combination of the desire of rationalisation of organic reactivity and of the need for ordering the experimental data both for elegance and for reciprocal understanding. As a consequence, already since first days, some transformations were identified as, for example, reductions, others as electrophilic substitutions, etc. This kind of classification is very crude, most of the time disregarding mechanism, but it showed large use by the chemist community because it sounds familiar and is sufficiently clear to permit an immediate understanding of the area of reactivity people are debating.

Obviously such a general classification scheme was revealed insufficient as soon as the requests of reaction storage and retrieval left the field of private communication to go towards the management of giant reaction databases. Here, there is a double purpose; on the one hand, the data must be as precise as possible in order to give the correct answer to a narrow query; on the other hand, they must be as general as possible so to give dynamic reaction classes and, consequently, permit the most comprehensive answer to a large query. Thus, many algorithms were developed with the aim of optimal handling of reaction classification.[2]

A second large area of application of reaction classes is represented by the studies on reaction prediction (or more generally reactivity prediction). Here, an optimal classification system is not an absolute requirement; nevertheless, both the management of the reaction prediction machine and the need of communicating the results often forced the developers to face the problem.[3-8]

A fast run through the literature gives the unexpected impression of the existence of several nice and efficient systems for reaction classification. For example, each database producer suggests at least one solution to the problem with emphasis on efficiency or on completeness, etc. However, as soon as the interested reader goes deeper into the offered solutions he/she rapidly realises that the approach differences seldom affect the main principles adopted. It is possible to easily identify two independent ways of classification: the first based on substructure definition of reaction centres; the second using physical descriptors. It is also possible to go a step further remembering the words of Lawson: "*.... a reaction is the optimisation of achieving an effect or property, although this is generally masked by the use of structural descriptors as a synonym for properties.*";[9] this states the real meaning of the concept "reaction", consequently giving a unitary view of the classification problem.

Because the interest of our study is directed to the reaction classification with the aim of reactivity analysis we will shortly discuss only those approaches concerned with the same aim. There are three main aspects to consider: a) the object of the classification (either structure or property); b) the level of generalisation of the classification; c) the choice of a hierarchical (stepwise) or fixed classification.

The structural representation of reactions is certainly the most common and widespread approach. The historical usage that chemists make of structural graphs to represent all the chemical facts has so accurately covered the field that every attempt to leave it out has been contested and would still encounter many difficulties. Thus the only, to our knowledge, reported classification by property[10] had to be diluted inside a "classical" classification by structure. Generalisation is the most important and, at the same time, subjective aspect of the classification. In fact, two conflicting forces act on it: large generalisation means comprehensive but inaccurate classification; restricted generalisation gives the opposite result. It must be absolutely clear that classification is impossible without generalisation.[11-13] A solution to circumvent the problem of generalisation is to adopt a hierarchical system.[14] This way, it is possible to have a gradual increase of generalisation that permits a dynamic choice of the analysis depth.

The final argument of this introduction considers the role of similarity in reaction classification. It is obvious that if we wish to subdivide data (in our case reactions) into sets we, willingly or not, are thinking in terms of similarity. Therefore its use is always guaranteed. But when implicitly used, there is an important disadvantage: inside one set, it is impossible to assure that if A is similar to B and B is similar to C, then also A and C are similar because the property that makes A similar to B can be different from the property that makes B similar to C. Nevertheless similarity is seldom explicitly used for reaction classification, but it is often considered when defining a reaction search system. In fact, many examples of explicit references to similarity as the property for inclusion of a reaction in a search set are known.[15]

We have already mentioned that our classification system has the three following features: is a property based method; is hierarchical; it uses similarity. We would like to argue about the choice of these options. Our personal experience in the fields of organic synthesis planning and of reaction prediction has a main objective: put a rationale into organic chemistry which is general and reliable enough to help solving known and unknown synthetic problems. Generalisation is achieved by the continuous search for descriptors that, though dependent on structures, represent atomic or molecular properties. In fact, if the property is quantitative, we can directly view the generalisation level we are using. Hierarchy is the tool to manage generalisation; it allows to go up and down the generalisation levels without need of data re-elaboration. This makes the classification more solid. Similarity is referred to explicitly, thus eliminating all the misinterpretations concerning the features we are comparing. Similarity has the power of increasing both the generalisation and the reliability.

Descriptors

The choice of the descriptors that will be used to classify the reactions is determining the true connection with the molecular reactivity. It is thus of fundamental importance to consider explicitly what each descriptor is describing. In addition we must decide at this phase of the project the precision we would like to use in the classification. Unfortunately chemical reactivity is influenced by so many different factors that its complete description is out of question. We have to select a limited number of descriptors and realise a system that consistently uses them.

It is generally accepted that electronic factors as well as geometric features influence the molecular behaviour. In addition we could consider solvent effects, temperature, salts, etc. Nevertheless, when classifying reactions it is possible to restrict the description to fit personal needs. Remembering Lawson's statement, we selected two atomic properties as reaction descriptors: electronic energy and chemical potential.[16] These two descriptors are correlated, as known, by the following equation:

$$\mu = \partial E / \partial n$$

where μ is the chemical potential, E is the electronic energy and n is the number of electrons. Yet, it is possible to distinguish the role of the two descriptors; electronic energy represents the status of an atom dipped into a well-defined environment, chemical potential is the resistance that an atom opposes to the environment changes. A reaction can be correctly represented by the combination of the status (electronic energy) of its atoms and of their "willingness" to change or maintain it. Geometric factors are presently absent in our reaction description, but their addition could be straightforward as soon as a general geometric descriptor is defined.

In order to make everything clear let us remember our method of energy and potential calculation. In our model, molecules are described as sets of atoms that reach an equilibrium by distributing electrons until all the chemical potentials are equalised. Thus, the following equations control the electron flow:

$$\mu = -\chi = -k_1 Z_{star} Z_{star'} / (Z^0{}_{star'} R^0{}_{cov}) + k_2 \qquad \text{where:}$$

$$Z_{star} = Z - \sigma = Z - (N_1 + 0.85N_2 + 0.35(N_3 - 1)) \qquad \text{and}$$

$$Z_{star'} = Z - (N_1 + 0.85N_2 + 0.35N_3)$$

where χ is the atom electronegativity, Z is the atomic nuclear charge, σ is the Slater's core screening factor, N_1 is the number of the inner-shell electrons, N_2 is the number of the medium-shell electrons, N_3 is the number of the outer-shell electrons that changes, even fractionally, depending on the electrons moved from one atom to one of its neighbours.

$$Qtot = \Sigma_n Q_i$$

$$Q_i = q * (1 - \exp(-1/4 * (\chi_i - \chi_k)^2))$$

where q is the electron charge, Q_i is the atomic charge moved along i-k bond, Qtot is the atomic total charge. This equation can be further generalised in order to consider differences in electron flows related to different environments, where it becomes

$$Q_i = q * (1 - \exp(t * (\chi_i - \chi_k)^2))$$

where t is a parameter depending on the i-k bond order.

Once equilibrium is achieved, we can also find a measure of the importance of each atom in maintaining the equilibrium itself. This measure is the electronic energy difference between the complete structure and a hypothetical molecule where each atom is removed in turns.[17] Electronic energy is calculated by the following equation, and the atom importance by the difference between the corresponding energy sums over all atoms.

$$E = \Sigma_i E_i$$

$$E_i = k_3 * (A + B + C) - k_2 * N_3$$

$$k_3 = -k_1/(Z^0_{star'} * R^0_{cov})$$

where k_1 and k_2 are constants depending on the atom type; $Z^0_{star'}$ is the effective nuclear charge of the isolated atom for a complete electronic shielding; R^0_{cov} is the atomic covalent radius of the isolated atom.

$$A = (N^2 + aN - 2NN_1 - 2bNN_2 + N_1^2 + 2bN_1N_2 - aN_1 + b^2N_2^2 - abN_2) * N_3$$

$$B = 0.5 * (-2aN + 2aN_1 + 2abN_2 - a^2) * N_3^2$$

$$C = (a^{2/3}) * N_3^3$$

where a and b are Slater's coefficients: N is the atomic number; N_i are the shell occupation numbers.

$$W_k = | \Sigma_{1,n} E_i - (\Sigma_{1,k-1} E_i^{-k} + \Sigma_{k-1,n} E_i^{-k} + E^0_k) |$$

where E_i^{-k} is the energy of atom (i) and E^0_k is the energy of atom (k) in the molecule with atom (k) isolated from the molecule.

When making a reaction we formally operate a heavy change in the molecular environment, often making and breaking bonds and repositioning atoms in space. This change affects both the energy that directly depends on the atomic status, and the potential. The energy variation gives a measure of the perturbation suffered by the atom, but, being static in nature, it could happen that the final position is not too dissimilar with respect to the starting situation. On the contrary, the chemical potential is more sensitive and more exactly reflects the change. In conclusion we can be sufficiently sure that our descriptors are representative of reactions.

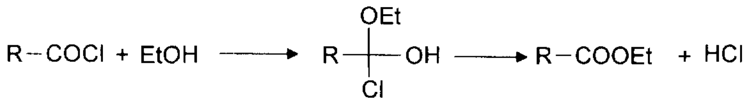
The reaction and its mechanism

We must take into consideration a different but important aspect that can influence our classification scheme. In fact, because we are going to use property descriptors, it is important to realise that they are completely determined by the studied molecule. As a consequence we have the descriptors of the educts and those of the products, but, in principle, we ignore the descriptors of every species that is between educts and products. For example, we can describe the following esterification:

$$\text{R–COCl} + \text{EtOH} \rightarrow \text{R–CO–OEt} + \text{HCl}$$

Scheme 1

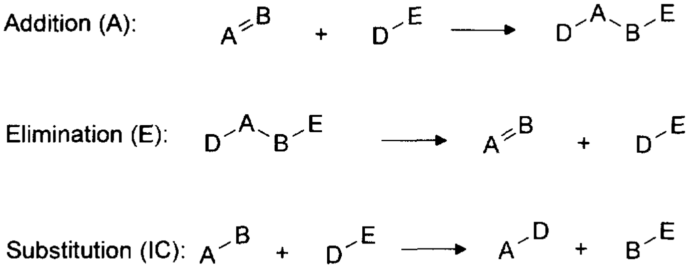
as a substitution reaction, where the chlorine atom is substituted by the alcohol residue. Or, the same reaction can be seen as a two step addition-elimination reaction:



Scheme 2

The difference of the two descriptions is so great that in all classification schemes they would be inserted in different clusters. It is thus of fundamental importance to adopt a standard method of reaction analysis. On the contrary, it is absolutely negligible, in this context, to ascertain what is the "real" mechanism. However, a good classification should "understand" that the reaction sketched in Scheme 1 is, in a different notation, equivalent to the two reactions of Scheme 2. This result is possible only using property descriptors. In addition, when wishing to go further into the details of a reaction, the system should permit to classify each single step.

Looking at the values of our descriptors when applied to different reactions, we arrived at the conclusion that all the reactions can be partitioned into only three fundamental classes: additions, eliminations, and substitutions, as shown in Scheme 3.



Scheme 3

For example, reductions are very often additions, whilst oxidations are very often eliminations; rearrangements are combinations of additions and eliminations, or of substitutions; etc. Moreover, each member of the reaction equation can participate to a different class, e.g. one compound is involved in an addition, the other in a substitution. Nevertheless the net result of a reaction is the aspect that we must consider for classification and, almost always, this aspect coincides with the greater change in the descriptor values. We will have a better perspective of this assertion after discussing the results.

This scheme apparently neglects isomerizations (e.g. rearrangements, racemizations, metathesis, etc.) but, as we will see in the Results section, all these reactions are more or less complex combinations of the three main reaction classes; in addition, the hierarchical description should allow for their correct identification. For the sake of completness, it should be mentioned that, in very special cases, the classification of complex transformations could be not descriptive enough, particularly when the starting and the final electronic states of the atoms remain more or less unchanged.

Similarity and reaction classes

We would like to discuss one more argument that arrives last but that is determining the efficiency of the entire system. Even with some good descriptors available, the classification is not finished. In fact an important phase still remains to be defined. Dividing objects into classes using calculated numerical values (that are continuous measures) must partition them into discrete dominions; consequently it becomes necessary to fix some thresholds. However, there is an inherent inaccuracy when we operate such an a priori fixed "black and white" choice because it can easily happen that an object outside of the threshold is much nearer to another object inside the threshold than to any other object in its own class. A possible solution to this problem is to determine a similarity measure that can correctly estimate the proximity of the objects. A second way could consider the transformation of the numerical values into discrete quantities, followed by the very easy division into classes and by a passage back to continuous values inside each class to order the objects. The second solution can sound like a trick to hide the "hard" separation mentioned a few lines above, an impression that is only partially true. In fact, using different measures to operate the discrete division, we "dilute" the importance of the single thresholds and, more important, we make possible the choice of the best classification at the usage stage. In this perspective the similarity between reactions belonging to the same class is determined by different operations, but it still remains linked to a numerical quantity that permits the attainment of an order.

Thus we conceived the following articulated system.

The first classification is done using the electronic energy (EE) changes of atoms in educts and products. As multiply bonded atoms are energetically rich, every addition that diminishes the bond order decreases the EE (and every elimination increases it); on the contrary, substitutions do not change atomic EE appreciably. We can thus define a threshold (positive for additions and negative for eliminations) that identifies the atoms involved in different transformations. In addition, a check is done on the integrity of the operation; i.e. for additions and eliminations it is important to verify that pairs of atoms participate in the transformation, eliminating those atoms that, even if energetically highly perturbed, are not concerned by this class of reactions. Then the surviving atoms are inserted into class A and E, respectively. On the contrary, all substitutions are put into class IC. We consequently arrive at three main classes defined by the EE.

From this point onwards, the property used for further classification is the chemical potential. This also changes differently depending on the type of the reaction. We can define four types of changes: for additions, for eliminations, for additive substitutions, and for eliminative substitutions. They are selected by the μ value (compared with a threshold) and by its sign (positive for additions and additive substitutions, negative for eliminations and eliminative substitutions). For each reaction it is therefore possible to locate and classify all the atoms that suffer a significant perturbation. The second level classes are determined just by the number of reacting atoms. A third level classification is obtained by the atomic classes determined as indicated previously. At the end of this analysis we have a hierarchical classification on three levels; we can finally order by total μ variation the reaction inside the third level classes. The ordering can obviously be applied at any level, representing the usage stage option referred a few lines above.

RESULTS AND DISCUSSION

Table 1. Set of Reactions to be Classified

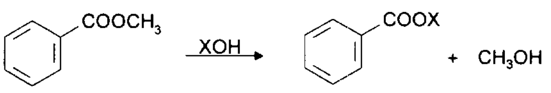
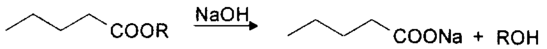
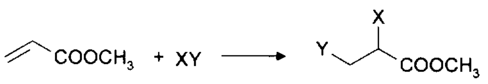
	X = H	R1
	X = Li	R2
	X = Na	R3
	X = K	R4
	R = CH$_3$	R5
	R = C_2H_5	R6
	R = Ph	R7
	R = CF$_3$	R8
	XY = HOH	R9
	XY = HNH$_2$	R10
	XY = HCl	R11
	XY = HCH$_3$	R12
	XY = HCH$_2$COCH$_3$	R13

Table 1. Set of Reactions to be Classified. Continued

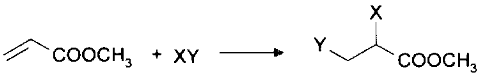
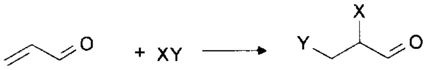
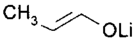
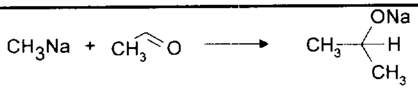
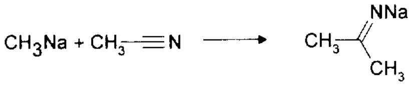
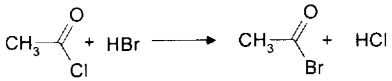
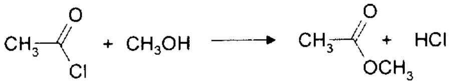
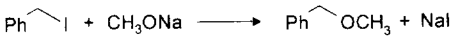
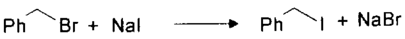
	XY = NaOH	R14
	XY = NaNH$_2$	R15
	XY = NaCl	R16
	XY = NaCH$_3$	R17
	XY = NaCH$_2$COCH$_3$	R18
	XY = NaCH$_3$	R19
	XY = NaBr	R20
	XY = 	R21
	XY = HNHCH$_3$	R22
	XY = HOCH$_3$	R23
$CH_3{-}X$ + NaY $\longrightarrow$ $CH_3{-}Y$ + NaX	X = F, Y = I	R24
	X = Cl, Y = I	R25
	X = OAc, Y = I	R26
	X = OMs, Y = I	R27
	X = OTs, Y = I	R28
$CH_3{-}X$ + NaY $\longrightarrow$ $CH_3{-}Y$ + NaX	X = F, Y = CN	R29
	X = Cl, Y = CN	R30
	X = OMs, Y = CN	R31
	X = OTs, Y = CN	R32
		R33
		R34
		R35
		R36
		R37
		R38
$CH_3{-}I$ + CH_3ONa $\longrightarrow$ CH_3OCH_3 + NaI		R39
$CH_3{-}I$ + NaBr $\longrightarrow$ $CH_3{-}Br$ + NaI		R40

Table 1. Set of Reactions to be Classified. Continued

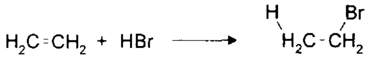
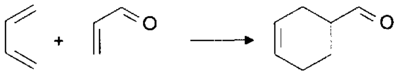
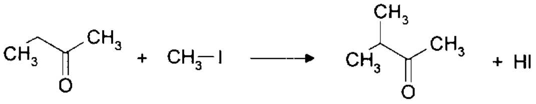
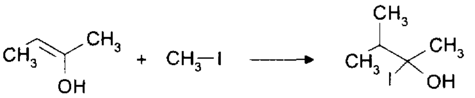
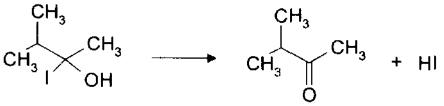
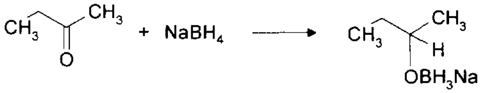
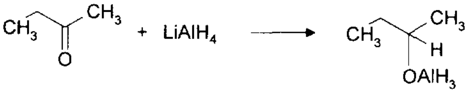
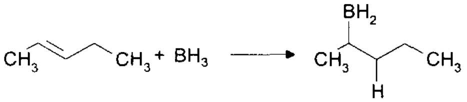
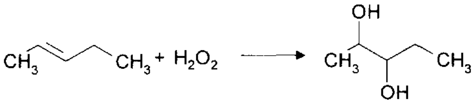
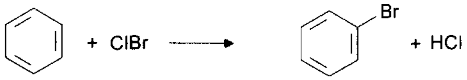
$HC\equiv CH + HBr \longrightarrow H{-}HC{=}CH{-}Br$	R41
	R42
	R43
	R44
	R45
	R46
	R47
	R48
	R49
	R50
	R51

Table 1. Set of Reactions to be Classified. Continued

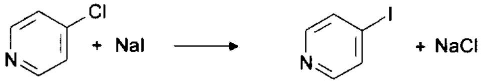
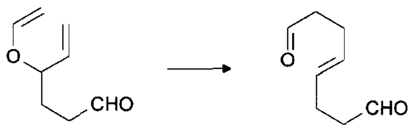
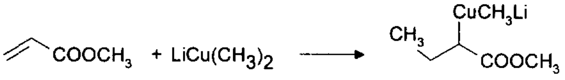
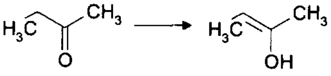
	R52
	R53
	R54
	R55

Table 2. Classification of Reactions of Table 1.

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
43	A	7	4A/3SA	8.31
54	A	6	2A/3SA/1SE	4.21
21	A	6	3A/2E/1SA	0.14
9	A	5	2A/1SA/2SE	4.30
23	A	5	2A/1SA/2SE	3.54
34	A	4	2A/2SA	5.77
48	A	4	2A/2SA	5.53
33	A	4	2A/2SA	5.49
47	A	4	2A/2SA	4.74
11	A	4	2A/2SA	4.30
50	A	4	2A/2SE	4.14
20	A	4	2A/2SE	2.81
41	A	4	2A/1SA/1SE	4.63
14	A	4	2A/1SA/1SE	4.43
10	A	4	2A/1SA/1SE	4.21
22	A	4	2A/1SA/1SE	4.17
15	A	3	2A/1SA	4.29
13	A	3	2A/1SA	4.16
12	A	3	2A/1SA	4.16

[a] Main reaction class; calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. [c] Classification based on the perturbation type. [d] Order obtained using chemical potential variations.

Table 2. Classification of Reactions of Table 1. Continued

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
19	A	3	2A/1SA	4.12
17	A	3	2A/1SA	4.10
18	A	3	2A/1SA	4.08
45	A	3	2A/1SA	4.05
49	A	3	2A/1SA	3.97
16	A	3	2A/1SA	3.66
42	A	3	2A/1SE	3.96
46	E	4	2E/1SA/1SE	-4.60
53	IC	7	2A/2E/1SA/2SE	-1.27
8	IC	6	4SA/2SE	0.46
32	IC	6	4SA/2SE	0.28
37	IC	5	5SA	1.78
30	IC	5	4SA/1SE	1.66
29	IC	5	4SA/1SE	0.86
31	IC	5	4SA/1SE	0.33
55	IC	4	1A/1E/1SA/1SE	0.97
39	IC	4	2SA/2SE	1.23
2	IC	4	2SA/2SE	0.06
7	IC	4	2SA/2SE	-0.03
24	IC	4	2SA/2SE	-1.07
26	IC	4	1SA/3SE	-1.29
35	IC	3	3SE	-0.23
1	IC	3	1SA/2SE	-0.07
6	IC	3	1SA/2SE	-0.09
3	IC	3	1SA/2SE	-0.09
4	IC	3	1SA/2SE	-0.1
36	IC	3	1SA/2SE	-0.12
5	IC	3	1SA/2SE	-0.14
52	IC	3	1SA/2SE	-0.30
25	IC	3	1SA/2SE	-0.33
28	IC	3	1SA/2SE	-1.22
27	IC	3	1SA/2SE	-1.23
44	IC	3	3SA	0.44
38	IC	2	1SA/1SE	0.04
40	IC	2	1SA/1SE	0.04
51	IC	2	2SE	-1.11

[a] Main reaction class; calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. [c] Classification based on the perturbation type. [d] Order obtained using chemical potential variations.

For the sake of understanding we will illustrate the system operations using a set of very simple examples planning the presentation of more complex examples in a subsequent section. We selected 55 reactions (see Table 1) that include additions (both electrophilic and nucleophilic), substitutions, one elimination only (because all the additions can be considered eliminations if looked at from right to left), and two rearrangements, one of which is an enolization (just to show that a rearrangement is the combination of addition and elimination steps). Table 2 reports the classification obtained by all the steps and Figure 1 shows a graphical representation of the class hierarchy. In the following we will deepen only some chosen examples.

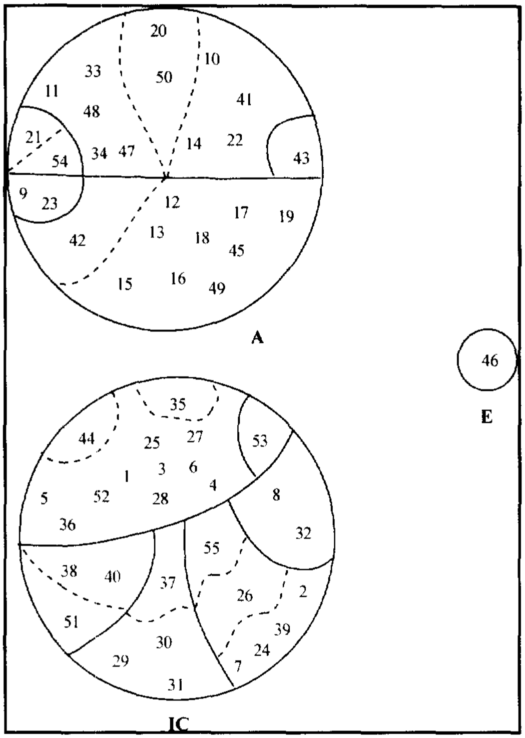


Fig. 1. Classification of reactions of Table 1. Addition, elimination, and substitution, areas contain the corresponding main classes. In each class circle, solid lines separate second level classes and dotted lines third level classes.

All additions are correctly selected by the energy descriptor, including R21 that is a mixed addition - rearrangement. R43, the single cycloaddition of the set, is well separated from the others. There are at least 2 atoms that are classified additive by μ and this generally determines a large positive values for the total μ descriptor. It is worth to note that this is not a strict requirement (e.g. R21). All 1,2 additions to carbon-hetero bonds fall in the same subclass (2A/2SA), as do all additions of carbon-heteroatom pairs to carbon-carbon double bonds (2A/1SA).

In the case of substitutions we can observe that: a) all the ester hydrolyses are grouped together by a larger number of SE atoms; the only exception is R8 that has a particular alcohol residue (OCF_3). A similar separation is present in halide substitution where all the uncertain reactions are grouped by SE ≥ SA atoms and $\sum\mu$ is negative or null, whilst the others show SA>SE atoms and their $\sum\mu$ is positive. Here too, there is an exception, R39, which is a substitution with SA=SE and with $\sum\mu$ clearly positive. R53 is a rearrangement that is energetically classified in the IC class; obviously looking at atom classification we find a 2A/2E case that correctly subclassifies R53.

As a last comment we would like to show that the descriptors correctly predict the result of a transformation despite of its mechanism. R44 (IC, 0.44) can be seen as the combination of R55 (IC, 0.97), R45 (A, 4.05), and R46 (E, -4.60), that gives a net result of IC, 0.42. This demonstrates the overall reliability of the system. In conclusion, we can affirm that the above simple examples show the possibility of reaction classification by this hierarchical system. It is worth remembering that, because the hierarchy allows the choice of the analysis level, it is always possible to get different hints just by changing the hierarchy depth.

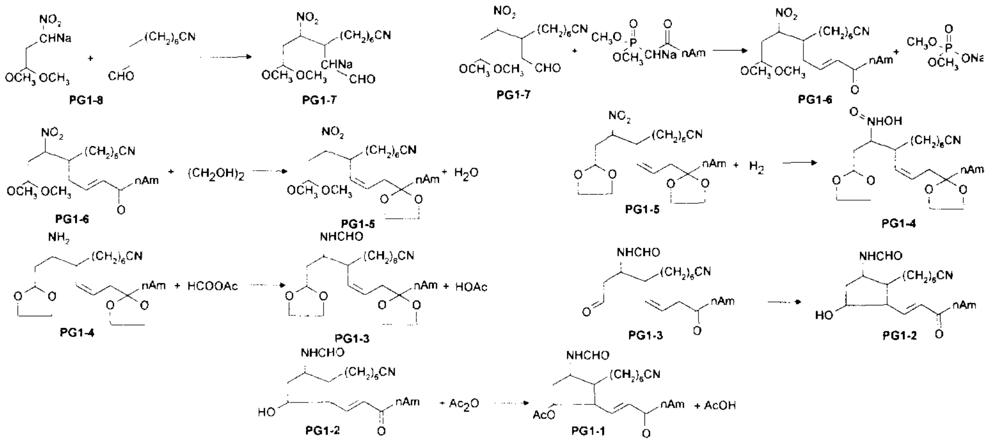


Fig. 2. Steps of the first synthesis of PGE2

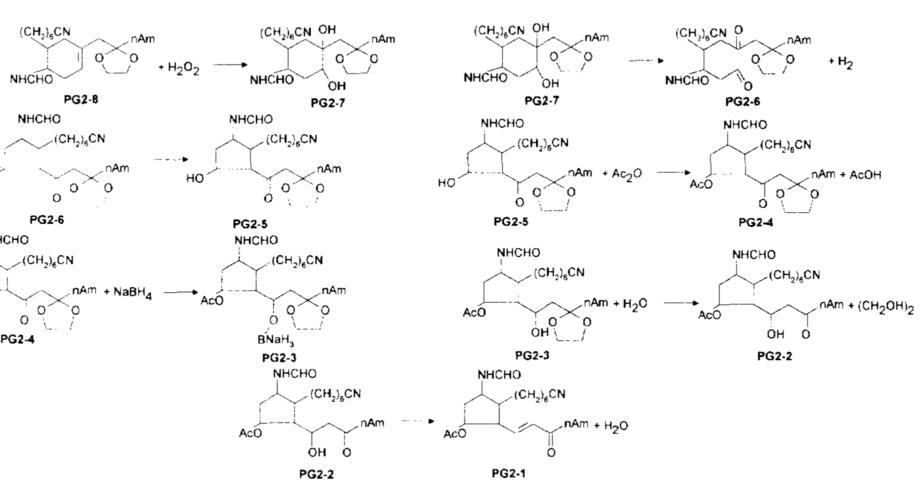


Fig. 3. Steps of the second synthesis of PGE2

Other examples

A further application of the classification system concerns a small set of prostaglandin derivatives (see Figures 2, 3, 4, 5) augmented by reactions taken from Methoxatin and Sirenin syntheses (Figures 6, 7).[18] The structural complexity of these compounds generates a greater number of subsets. It is less probable that the same number of atoms are affected in such different molecules. It is clear that the main classification (that based on energy) still puts into the correct class the reactions, generating the corresponding addition, elimination, and substitution sets. On the contrary, the division using both the number of the used atoms and their distribution are now far less diagnostic giving raise to too many and too assorted classes (Table 3).

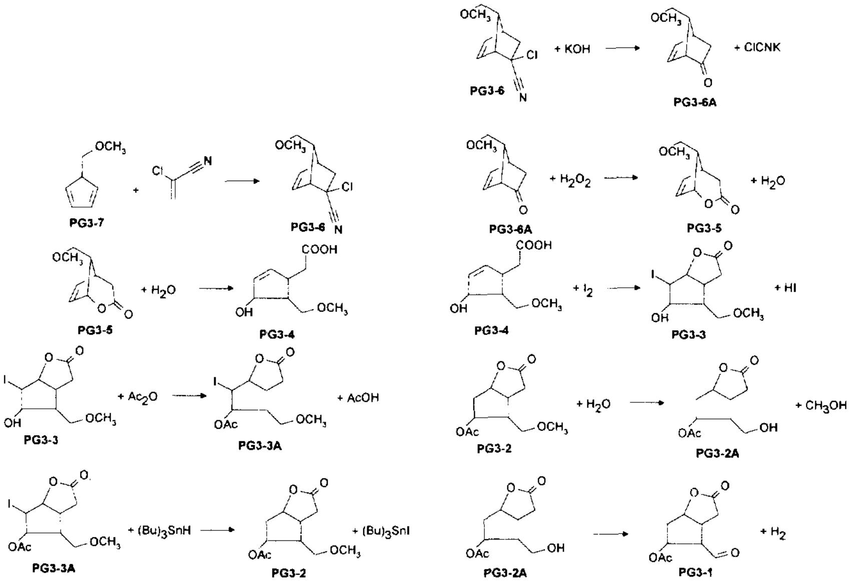


Fig. 4. Steps of the synthesis of one precursor of PGE2

Nevertheless, the flexibility of our system allows for a different view of the similarity measures. In fact we can slightly change the method and use a different basis to operate the analysis. Inside each main class we can locate and group together all those reaction that share the number of the most perturbed atoms; i.e. in reaction M3/M2 there are 7 affected atoms, 4 of which suffer an addition thus contributing more to the $\sum\mu$. Grouping the reactions in this way we obtain 9 subclasses only (Table 4), three for each main class. Let us examine these results.

Main class: additions. The first subclass contains all the cycloadditions with all values very similar. The second subclass contains all single additions to multiple bonds. This set is more articulated as far as the values are concerned (the range is 2.96 to 7.28). The highest values concern reactions that involve organometallic or similar reagents. Well separated are two reactions (PG3-6/PG3-6A, PG4-7A/PG4-6) that are not pure additions but involve one addition and one elimination at the same time.

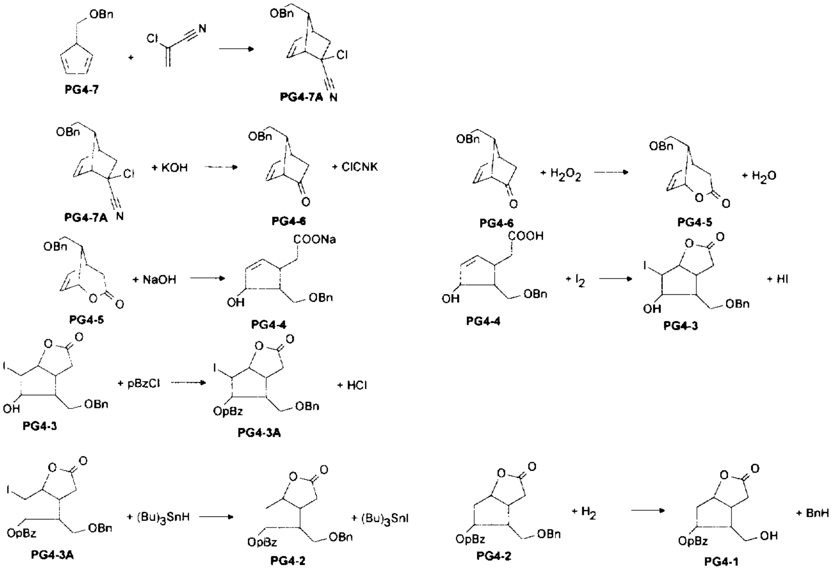


Fig. 5. Steps of the synthesis of another precursor of PGE2

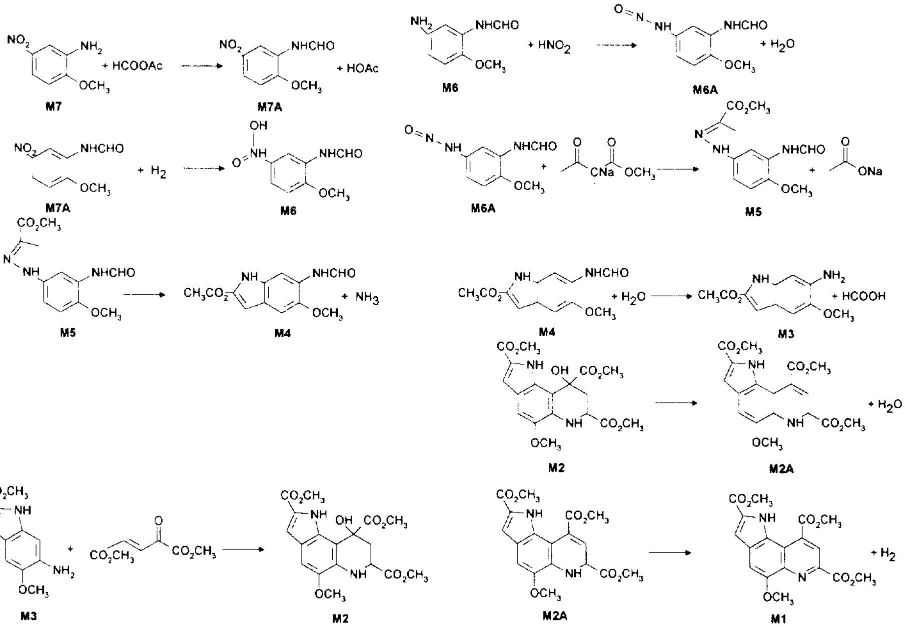


Fig. 6. Steps of the synthesis of Methoxatin

Main class: eliminations. The first subclass contains only one reaction; it is an isolated case of double elimination. The reaction clearly implies two separate steps but the representation chosen is in agreement with the classification. The second subclass includes true eliminations, where the elimination of a molecule of hydrogen ranks better than the elimination of water. The third subclass, on the contrary, includes reactions that, even describing an elimination step, are more complex because other transformations also occur.

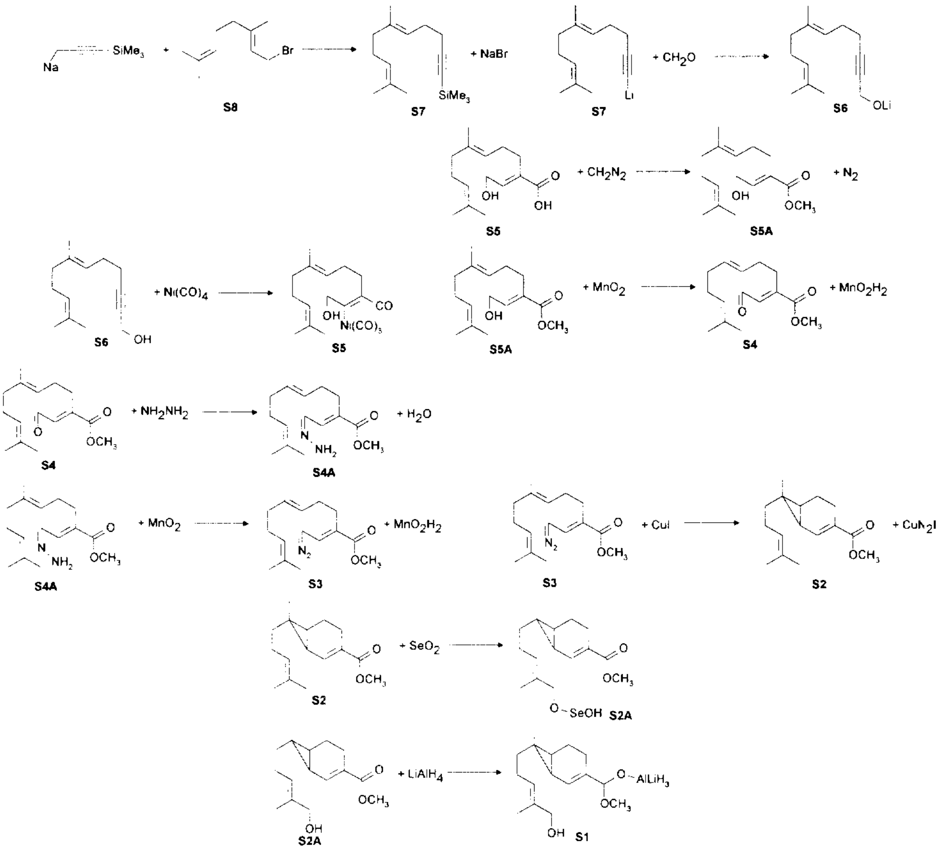


Fig. 7. Steps of the synthesis of Sirenin

Table 3. First Classification of Reactions of Figures 2-7

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
PG1-3/PG1-2	A	8	3A/1E/3SA/1SE	5.09
PG4-4/PG4-3	A	8	2A/2SA/4SE	3.39
PG4-7/PG4-7A	A	7	4A/3SA	8.30
M3/M2	A	7	4A/1SA/2SE	8.93
M7A/M6	A	7	2A/3SA/2SE	6.09
PG1-6/PG1-5	A	7	2A/1SA/4SE	4.86
PG3-4/PG3-3	A	7	2A/1SA/4SE	2.96

[a] Main reaction class: calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. [c] Classification based on the perturbation type. [d] Order obtained using chemical potential variations.

Table 3. First Classification of Reactions of Figures 2-7. Continued

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
S3/S2	A	6	4A/1SA/1SE	8.64
PG2-6/PG2-5	A	6	2A/4SA	5.75
PG1-5/PG1-4	A	6	2A/3SA/1SE	6.55
S2A/S1	A	6	2A/3SA/1SE	5.74
S6/S5	A	6	2A/2SA/2SE	3.99
PG3-6/PG3-6A	A	6	2A/2E/1SA/1SE	1.35
PG4-7A/PG4-6	A	6	2A/3E/1SE	-3.17
PG3-7/PG3-6	A	5	4A/1SA	8.16
S2/S2A	A	5	2A/3SA	7.28
PG2-4/PG2-3	A	5	2A/3SA	5.33
PG1-8/PG1-7	A	5	2A/3SA	4.07
PG2-8/PG2-7	A	5	2A/1SA/2SE	4.27
S7/S6	A	4	2A/2SA	5.23
M5/M4	E	13	1A/1E/3SA/8SE	-0.30
S4A/S3	E	8	1A/1E/4SA/2SE	-1.91
S5A/S4	E	7	1A/2E/2SA/2SE	-1.00
M2A/M1	E	6	2E/2SA/2SE	-4.21
PG2-3/PG2-2	E	6	2E/4SE	-5.40
PG2-7/PG2-6	E	6	4E/2SE	-10.84
M2/M2A	E	5	2E/3SE	-5.86
PG2-2/PG2-1	E	4	2E/1SA/1SE	-4.38
PG3-2A/PG3-1	E	4	2E/2SE	-6.11
PG1-7/PG1-6	IC	8	1A/1E/2SA/4SE	1.32
M6A/M5	IC	7	1A/1E/2SA/3SE	2.15
PG4-3/PG4-3A	IC	7	6SA/1SE	1.34
M6/M6A	IC	7	4SA/3SE	0.59
S4/S4A	IC	6	1A/1E/3SA/1SE	0.18
S5/S5A	IC	5	1A/1E/3SE	-2.91
PG4-6/PG4-5	IC	5	2E/2SA/1SE	-2.45
PG4-2/PG4-1	IC	5	4SA/1SE	1.49
PG3-3/PG3-3A	IC	5	4SA/1SE	0.88
PG4-3A/PG4-2	IC	5	3SA/2SE	0.37
M7/M7A	IC	5	3SA/2SE	0.10

[a] Main reaction class: calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. [c] Classification based on the perturbation type. [d] Order obtained using chemical potential variations.

Main class: substitutions. In this class there are not atoms contributing more to the result. Therefore the division has been done considering together all reactions involving elimination/addition atoms; all those involving elimination atoms; and all those remaining. The obtained classification evidently separates correctly similar reactions. In the first subclass there are those reactions that are particular cases of substitutions (e.g. the esterification of an acid with diazomethane). In the second subclass there are two Bayer-Villiger reactions. Finally, in the third subclass we can easily find all true substitutions. There is only one particular case, reaction PG4-5/PG4-4, that has an atom classified as elimination atom; this is the only misclassification present in all the classes, even if the main classification is correct.

Table 3. First Classification of Reactions of Figures 2-7. Continued

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
PG3-6A/PG3-5	IC	4	2E/2SA	-2.34
S8/S7	IC	4	4SA	1.66
PG1-4/PG1-3	IC	4	4SA	0.56
PG3-5/PG3-4	IC	4	3SA/1SE	0.48
M4/M3	IC	4	2SA/2SE	-0.10
PG2-5/PG2-4	IC	4	1SA/3SE	-0.54
PG4-5/PG4-4	IC	2	1E/1SE	-3.34
PG3-3A/PG3-2	IC	2	1SA/1SE	0.01
PG3-2/PG3-2A	IC	2	1SA/1SE	-0.002
PG1-2/PG1-1	IC	1	1SA	0.08

[a] Main reaction class; calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. [c] Classification based on the perturbation type. [d] Order obtained using chemical potential variations.

Table 4. Second Classification of Reactions of Figures 2-7

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
M3/M2	A	7	4A/1SA/2SE	8.93
S3/S2	A	6	4A/1SA/1SE	8.64
PG4-7/PG4-7A	A	7	4A/3SA	8.30
PG3-7/PG3-6	A	5	4A/1SA	8.16
S2/S2A	A	5	2A/3SA	7.28
PG1-5/PG1-4	A	6	2A/3SA/1SE	6.55
M7A/M6	A	7	2A/3SA/2SE	6.09
PG2-6/PG2-5	A	6	2A/4SA	5.75
S2A/S1	A	6	2A/3SA/1SE	5.74
PG2-4/PG2-3	A	5	2A/3SA	5.33
S7/S6	A	4	2A/2SA	5.23
PG1-3/PG1-2	A	8	3A/1E/3SA/1SE	5.09
PG1-6/PG1-5	A	7	2A/1SA/4SE	4.86
PG2-8/PG2-7	A	5	2A/1SA/2SE	4.27
PG1-8/PG1-7	A	5	2A/3SA	4.07
S6/S5	A	6	2A/2SA/2SE	3.99

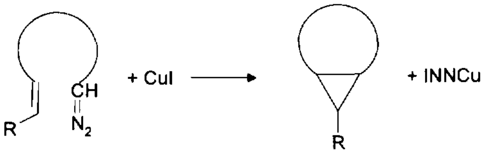
[a] Main reaction class; calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. Not used [c] Classification based on the atom main perturbation type. [d] Order obtained using chemical potential variations.

Table 4. Second Classification of Reactions of Figures 2-7. Continued

Reaction	Class I[a]	Class II[b]	Class III[c]	Order[d]
PG4-4/PG4-3	A	8	2A/2SA/4SE	3.39
PG3-4/PG3-3	A	7	2A/1SA/4SE	2.96
PG3-6/PG3-6A	A	6	2A/2E/1SA/1SE	1.35
PG4-7A/PG4-6	A	6	2A/3E/1SE	-3.17
PG2-7/PG2-6	E	6	4E/2SE	-10.84
PG3-2A/PG3-1	E	4	2E/2SE	-6.11
M2/M2A	E	5	2E/3SE	-5.86
PG2-3/PG2-2	E	6	2E/4SE	-5.40
PG2-2/PG2-1	E	4	2E/1SA/1SE	-4.38
M2A/M1	E	6	2E/2SA/2SE	-4.21
S4A/S3	E	8	1A/1E/4SA/2SE	-1.91
(S4A/S3	E	8	3A/2E/1SA/2SE	0.03)
S5A/S4	E	7	1A/2E/2SA/2SE	-1.00
(S5A/S4	E	6	2A/2E/2SE	0.32)
M5/M4	E	13	1A/1E/3SA/8SE	-0.30
M6A/M5	IC	7	1A/1E/2SA/3SE	2.15
PG1-7/PG1-6	IC	8	1A/1E/2SA/4SE	1.32
S4/S4A	IC	6	1A/1E/3SA/1SE	0.18
S5/S5A	IC	5	1A/1E/3SE	-2.91
PG4-6/PG4-5	IC	5	2E/2SA/1SE	-2.45
PG3-6A/PG3-5	IC	4	2E/2SA	-2.34
S8/S7	IC	4	4SA	1.66
PG4-2/PG4-1	IC	5	4SA/1SE	1.49
PG4-3/PG4-3A	IC	7	6SA/1SE	1.34
PG3-3/PG3-3A	IC	5	4SA/1SE	0.88
M6/M6A	IC	7	4SA/3SE	0.59
PG1-4/PG1-3	IC	4	4SA	0.56
PG3-5/PG3-4	IC	4	3SA/1SE	0.48
PG4-3A/PG4-2	IC	5	3SA/2SE	0.37
M7/M7A	IC	5	3SA/2SE	0.10
PG1-2/PG1-1	IC	1	1SA	0.08
PG3-3A/PG3-2	IC	2	1SA/1SE	0.01
PG3-2/PG3-2A	IC	2	1SA/1SE	-0.002
M4/M3	IC	4	2SA/2SE	-0.10
PG2-5/PG2-4	IC	4	1SA/3SE	-0.54
PG4-5/PG4-4	IC	2	1E/1SE	-3.34

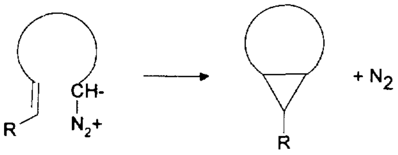
[a] Main reaction class; calculated using electronic energy changes. [b] Classification based on the number of perturbed atoms. Not used [c] Classification based on the atom main perturbation type. [d] Order obtained using chemical potential variations.

Specific examples. S3/S2 is a typical example of 1+2 cycloaddition; thus, its classification in the group of 2+4 cycloadditions is surprising, where it is clear that the atoms involved in the addition are four. This result is determined by the mechanism envisaged for the reaction (Scheme 4):



Scheme 4

An alternative mechanism can be the following (Scheme 5):



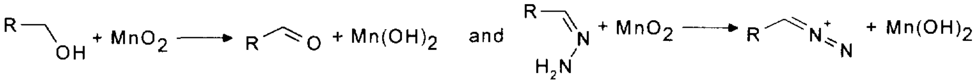
Scheme 5

The calculation, in this last case, gives the classification: Addition, 6, 3A/1E/2SA, 2.23. The two possibilities can be merged adding another step to Mech1: I-N=N-Cu → CuI + N_2, classified as: Elimination, 2, 2E, -6.41; whose combination with Mech1 gives: Addition, 6, 3A/1E/1SA/1SE, 2.23, very similar to Mech2. The true mechanism is clearly different involving some redox steps, but it is difficult to be represented. Nevertheless, it is important to emphasise that the classification scheme correctly determines the main class (Addition) and gives the expected correlation through mechanism combination.

S2/S2A is another example of apparent misclassification. In fact, it is an insertion reaction and not an addition. But, here again, it is difficult to represent the true mechanism; the chosen representation implies an addition to the SeO_2 molecule and the classification is its consequence; the successive hydrolytic step cannot modify the overall classification. The result is in agreement with PG2-8/PG2-7 reaction that is also an oxidative addition.

PG1-8/PG1-7 is a Michael addition and both the classification and the calculated $\Sigma\mu$ value are in good agreement with the simple examples reported in Table 1 (R9-R19, R22, R23, R45), particularly with R18 that is a very similar reaction.

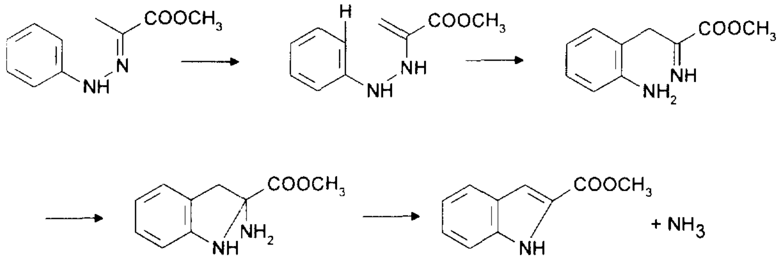
S5A/S4 and S4A/S3 are two very similar oxidations. Here again, it is possible to envisage a different mechanism (Scheme 6):



Scheme 6

The main classification does not change, but the calculated values and the subclassifications are affected. The true mechanism involves some electron shifts and any representation is acceptable; nevertheless, the similarity between the two reactions is confirmed.

M5/M4 is a Fischer indole synthesis that involves four steps, at least (Scheme 7):



Scheme 7

The one-step representation that we chose can only give a rough idea of the complete transformation. It is interesting that the great number of affected atoms (13) suggest a complex mechanism in agreement with the experimental data.

PG4-3/PG4-3A and PG4-2/PG4-1 are the protection - deprotection steps of the hydroxyl group. They are classified together and show very similar $\sum\mu$ values, despite of the fact that the PG4-2/PG4-1 is a reductive substitution.

Also M6/M6A, PG1-4/PG1-3, and M7/M7A, reactions are classified together and show very similar $\sum\mu$ values in agreement with the similar participating groups, NH_2 and an electron withdrawing partner (either NO or CHO).

The results shown above clearly demonstrate the validity of our classification scheme. Undoubtely, the subdivision into classes reflects the nature of the reactions. Nevertheless, a few general comments can be helpful.

First, it is important to bear in mind that any classification scheme is, willingly or not, dependent on the reaction description. This appears clear looking at our examples. Just to mention one, consider reactions PG3-6/PG3-6A and PG4-7A/PG4-6; they are, as far as the principal compound is concerned, eliminations, but the choice to include into the representation also the corresponding formal addition of KOH onto the CN substructure throws them into the addition class, in complete agreement with the classification scheme. However, a different description, e.g. neglecting the small molecule destiny, would have given a different result. Also the location of the most representative mechanistic step is determinant; for example the ester reduction of S2A into S1, involves at least three steps: an addition to the ester group, the successive elimination to an aldehyde, and the last addition to the carbonyl group. It is clear that the net result is an addition (two additions minus one elimination), but it is also clear that the classification would be different.

A second note concerns the choice of dividing all reactions into three main classes, only. The usual division into classes represented by the change incurred by the so-called "substrate", ignoring the other reaction component, is still a possibility. It is a correct and well understood classification. But it is too rigid and not general enough. Our scheme can put together reactions that, even if formally different (e.g. hydride reduction and hydrogen peroxide addition), belong to the same class.

Third, the subclassification obtained by using the type and number of changed atoms gives a further improvement to the general scheme; it groups reactions that share more than the main mechanism. In addition, a choice of the procedure used to obtain the subclassification permits a different level of precision and presents a highly interesting flexibility as shown by our two sets of examples.

Finally, it remains to comment on the numerical ordering given by the $\sum\mu$. This factor has not been fully considered in the present paper because it does not strictly concern reaction classification. Nevertheless, it is important to remember that the possibility of ordering objects inside similarity sets could be the very solution to many problems, including reactivity and synthesis design.

Conclusion

The classification scheme reported in the present paper uses similarity as the tool of the reaction allocation. The hierarchical structure and the use of calculated descriptors permit a flexible classification together with the possibility of ordering objects inside each class. The application of this procedure has shown that it is possible to include reactions in the expected class, but it has also shown the need of fixing the representation style of each reaction. In conclusion the obtained scheme is very nice for application in the synthesis planning and reaction prediction fields, but in order to be applied to data management it needs the definition of a general representation style.

Acknowledgement

Partial financial support by the Consiglio Nazionale delle Ricerche, and by the Ministero dell'Universita' e della Ricerca Scientifica e Tecnologica, is gratefully acknowledged. The authors would like to thank the Organising Committee of the "7th International Conference on Mathematical Chemistry" for the invitation to present part of this work.

REFERENCES AND NOTES

1. E.g. the compendia: *Theilheimer, Houben-Weyl, Organic Reactions, Comprehensive Organic Transformations.*
2. E.G. the databases: REACCS, CASREACT, SYNLIB, ORAC, CROSSFIRE plus REACTIONS.
3. Gasteiger, J.; Hutchings, M.G.;Christoph, B.; Gann, L.; Hiller, C.; Low, P.; Marsili, M.; Saller, H.; Yuki, K. *Top. Curr. Chem.* **1987**, *137*, 19.
4. Zefirov, N.S.; Tratch, S.S. *Anal. Chim. Acta* **1990**, *235*, 115.
5. Ugi, I.; Bauer, J.; Blomberger, C.; Brandt, J.; Dietz, A.; Fontain, E.; Gruber, B.; v. Scholley-Pfab, A.; Senff, A.; Stein, N. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*, 3-16.
6. Jorgensen, W.L.; Laird, E.R.; Gushurst, A.J.; Fleischer, J.M.;Gothe, S.A.; Helson, H.E.; Paderes, G.D.; Sinclair, S. *Pure Appl. Chem.* **1990**, *62*, 1921.
7. Sello, G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 713.
8. Satoh, H.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 173.
9. Lawson, A.J. *J. Chem. Inf. Comput. Sci.***1992**, *32*, 675.
10. Rose, J.R.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74.
11. Hendrickson, J.B.; Sander, T. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 251.
12. Gasteiger, J.; Marsili, M.; Hutchings, M.G.; Saller, H.; Low, P.; Rose, P.; Rafeiner, K. *J.Chem.Inf. Comput. Sci.* **1990**, *30*, 467.
13. Blurock, E.S. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 505.
14. Ihlenfeldt, W.D.; Gasteiger, J. *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2613.
15. Blurock, E.S. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 607.
16. a) Sello, G. *Theochem* **1995**, *340*, 15.
    b) Baumer, L.; Sello, G. *J. Chem. Inf. Comput Sci.* **1992**, *32*, 125.
17. Sello, G. *J. Am. Chem. Soc.*, **1992**, *114*, 3306.
18. Corey, E.J.; Cheng, X. In *The Logic of Chemical Synthesis,* John Wiley & Sons, Inc.: New York, 1989; pp. 141, 165, 251, 253, 255, 258.

(*Received in UK* 13 *June* 1997; *revised* 6 *August* 1997; *accepted* 7 *August* 1997)