Similarity, diversity, and the comparison of molecular structures.

Guido Sello, Dipartimento di Chimica Organica e Industriale, Universita' degli Studi di Milano, via Venezian 21, 20133 Milano, Italia; e-mail:sellogui@icil64.cilea.it; phone:++39-022363469, Fax: ++39-022364369

Abstract

Similarity is a powerful tool for compound comparison and can be seen as a good method to predict toxicity. One similarity measure and its application to complex problem solving are described.

The importance of the determination of compound toxicity is a fundamental requisite for the introduction of new chemicals into daily use. However, the cost, in terms of both time and money, of an accurate experimental determination forbids an uncontrolled application of well established tests. In addition, the recent efforts oriented to decrease the number of animal tests because their cost and their relative reliability for human toxicity prediction have stimulated the research towards alternative approaches. (Polloth and Mangelsdorf 1997) Among these, theoretical predictions based on the correlation between a structure and its activity represent a powerful method for the selection of toxic candidates which can be then accurately tested by experiments. In order to assess the possibility of a correlation two requisites are needed: the availability of experimental toxicity data and a method for describing structures. Clearly, it is the second aspect that we are concerned with.

The description of chemical structures is implicitly a modelling activity. In fact, any representation of molecules must use a "pictorial" description of their structures. The difference is often related to the explicit or implicit way of performing the modelling activity. A second aspect that is fundamental is the understanding that there is not an universal mode of description; on the contrary, it is common to have different methods depending on the current application. In conclusion, we are going to introduce a particular method for describing structures and a system to compare structures, closing with a special attention to their potential application also to the toxicity field.

The Description System

The Molecular Descriptor

Many different structural features are used for the characterisation of molecules. Because we want to go as deep as possible we selected an atomic descriptor. (Baumer, L., and Sello, G. 1992) It must be a calculated quantity, independent of a particular molecular class, sensitive to atomic environment, and easy to understand. One such descriptor is atom electronic energy (AE). Generally speaking, AE depends on atom chemical potential following the relation: $AE = \int \mu \, dn$, where μ is the chemical potential and dn is the electron variation. Because both μ and dn depend on electron distribution, that is sensitive to atomic neighbourhood, we obtain a diverse AE for each diverse situation, both if it is a stable or an unstable molecular state. μ is calculated using the following equation:

$$\mu = -k_1 \times Z_{eff} / r + k_2$$

where k_1 and k_2 are constants that depend on atom type, r is the atomic covalent radii, and Z_{eff} is the nuclear effective charge. But

$Z_{eff} = N - (aN_1 + bN_2 + c(N_3-1))$, where a, b, and c, are Slater coefficients, N is the total number of electrons, N_1, N_2, and N_3, are the electronic occupation of the atomic shells; and

$r = Z'_{eff} / (Z'^{0}_{eff} \times r^0)$, where Z'_{eff} is the effective nuclear charge with complete electron shielding and r^0 is the standard covalent radii of the atom; then

$$\mu = -k_1 \times [N - (aN_1 + bN_2 + c(N_3-1))] \times [N - (aN_1 + bN_2 + cN_3)] / (Z'^{0}_{eff} \times r^0) + k_2$$

Consequently, it is possible to calculate AE using the following equation:

$$AE = k_3 \times (A + B + C) - k_2 \times N_3 + K, \text{ where}$$

$$A = (N_2 + aN - 2NN_1 - 2bNN_2 + N_1^2 + 2bN_1N_2 - aN_1 + b_2N_2^2 - abN_2) \times N_3$$

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

$$C = a^{2/3} \times N_3^3$$

Just by summing up all the atomic contribution we can calculate the electronic energy of a molecule: $E = \sum AE$. This is a quantity that is representative of a molecule in a well defined state. Therefore, if you change the molecular state, either electronically or geometrically, you will have different Es. AE and E can be seen as interesting molecular descriptors for all those situation where the electronic state is important. In principle, this is a fundamental atomic characteristic that can be used when it is important to evidentiate a molecular behaviour connected to atom interaction.

The Molecular Comparison

Another important issue is the use that is made of the descriptor. It is worth emphasising that the characteristics chosen to analyse molecules are not self-informative and their profitable use depends on the environment where they are inserted. One scenario is the goal of learning something about a compound by its comparison to other better known compounds. This has been a long lasting wish of chemists that has become more easy by the introduction of the modelling activity. In between the many experiences the use of the evaluation of structure similarity represents a recent development that has opened many new perspectives.

The assumption of a working definition of similarity is basilar to reach an agreement on the meaning of the

model. In fact, similarity can be seen as an attribute of an object with respect to a particular property; e.g. two objects can be defined similar because of their colour, or their taste, etc. It is thus impossible to have an absolute definition. In the case of molecules it is important to always keep clear the application area we are considering. For example, two molecules can have similar solubility in water, or they can have similar number of carbonyl groups, or similar geometry. In our system (Sello, G. 1992) we are comparing structures on an atom by atom basis, selecting atoms that have a similar relevance to the electronic state of each compound. In different words, we are more interested in the role that an atom has in a compound than in its absolute energy. To this end, we define the importance (weight) of an atom in a molecule, as

$$AW = |E_{tot} - (E_{tot-i} + E_i^0)|$$

where E_{tot} is the energy of the complete molecule, E_{tot-i} is the energy of a hypothetical molecule obtained from the original compound by eliminating the interactions of atom (i) with all the remaining atoms, and E_i^0 is the energy of the isolated atom (i). This represents a measure of the perturbation that atom (i) suffers because it is inserted in a particular neighbourhood. As a consequence, it is possible that the same atom type shows different weights depending on the molecule. (Leoni, B., and Sello, G. 1995) This measure is very sensitive; for example, if it is calculated in its most accurate form, it is sufficient a conformation change to change the AWs.

Once chosen the similarity measure it is worth to organise the structure comparison. This is not as banal as can be thought; the problem of structure orientation is always present, even in topological comparison. Among the alternatives we chose to develop a canonical comparison, i.e. we analyse the atoms of a molecule in a canonical order. (Sello, G., and Termini, M. 1996a) The canonicity is decided by the atom weights and by the atom - atom connections. Starting from the most representative atom (that with the highest weight) and following its bonds its neighbouring atoms are ordered by weights. The procedure is repeated until all the atoms are ordered. Then the method follows two different directions depending on the dimension of the comparison. Topological comparisons are made by matching atoms in the same positions in the canonical orders and selecting those atoms that have similar descriptor. The result is one or more chains of similar atoms (figure 1).

Spatial comparisons use the canonical orders in a different fashion. When we wish to match molecules also taking into account their geometry we have to solve the problem of their relative orientation in space. This is a well known and often discussed problem. In our view the atom weight has a fundamental role; thus, we use the canonical order determined by this characteristic also to orient the structures. In other words, we positioned one of the compound oriented using its three most important atoms (the first in

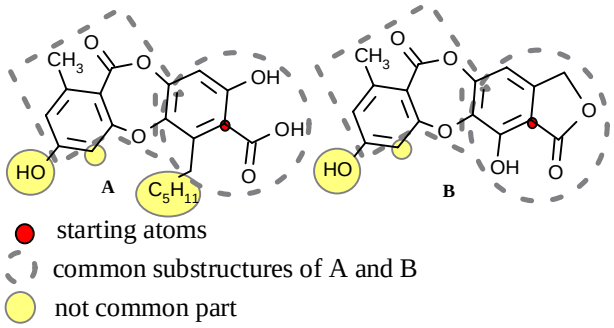

Figure 1. Similarity of variolaric and physodic acids.

the coordinate origin, the second along the X axis, the third in the positive Y part of the XY plane) and then we use all the possible relative orientations of the second compound given by its positioning using the same principle but with all the possible sequences of three consecutive atoms in its canonical order. Then, we select all the atom pairs that have similar AW and are sufficiently near in space. The best similarity result is chosen as the similarity of the two structures in space. In figure 2 is reported the result obtained by comparing the antitumoral compound Taxol with a modified natural compound. ((Sello, G., and Termini, M. 1996b) In this case bond flexibility is also considered.

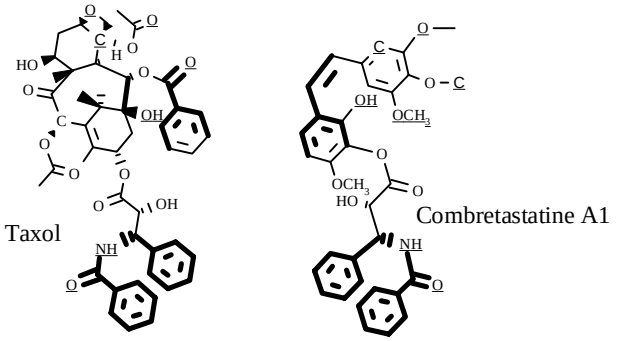

Figure 2. Similarity of Taxol and Combretastatine A1. Marked atoms are similar in weight and position.

The most evident difference is that also isolated atoms can become part of a similarity set regardless of their connections. However, the most important difference is that in the topological comparison it is possible to evidentiate a greater level of similarity but with less precision, whereas in the spatial comparison we accurately consider the atom position at the risk of missing some positive matching. It is obvious that the two alternatives have different applications.

Diversity, or the Similarity of Dissimilar Compounds

Once accepted the similarity concept, it is natural to consider the possibility of introducing its companion: molecular diversity. In principle, if it is possible to measure similarity it should be possible to measure diversity. This

so self-evident consequence should be carefully considered, because, if the intuitive assertion that two compounds are diverse can be obvious, the measure of their diversity can present many problems. First of all, similarity is an attribute of an object with respect to a feature; i.e. we must always define a relative similarity. For example we can affirm that two objects are similar in their colour. Can we affirm that two objects are diverse in their colour? Naturally, yes. But, can we measure their diversity? And order three objects using this measure? The problem has two faces. One is purely terminological; we use the term diversity but we mean scarcely similar. The other is much more important and concerns the reliability of a measure when it becomes very small, or, as is the case in our method, when it can be null because the measure is discrete. In fact, we measure similarity as the number of similar atoms; i.e. if a compound has no atoms similar to other two compounds we can affirm that they are diverse, but we cannot give a measure of this diversity.

Between many solutions presented in the literature (e.g. the use of many molecular descriptors to always guarantee a non-zero result) we propose a different approach. We assume that our similarity measure is a real measure, i.e. if one object A is distant X units from another object B that is distant Y units from a third object C then the distance between A and C is X+Y. It is obvious that if it is possible to measure a direct distance between A and C this will be different from X+Y, i.e. the indirect distance we are measuring is different from the direct distance. Nevertheless, if we measure all the distances between objects using the same standard they will be comparable. In this way we can have always non-zero similarity measures, even for highly dissimilar compounds.

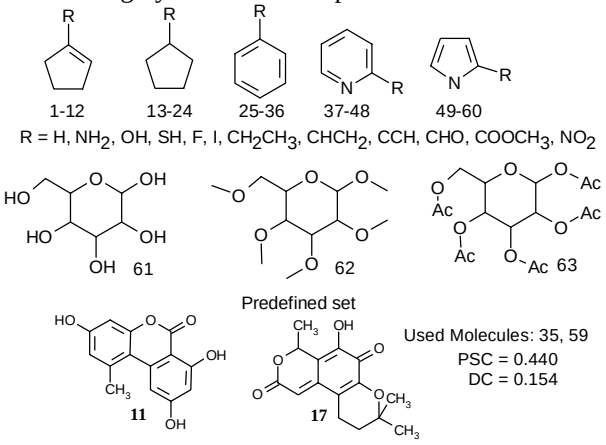

Figure 3. Molecule comparison using a predefined set

We developed two systems for measuring this new similarities, both of them use the same descriptor, atom weight, but the first compares two molecules through a predefined set of standard molecules, (Sello, G. 1998a), whereas the second slowly modifies the two molecules until a significative similarity measure is possible and weights the work required to make the modifications. In figures 3 and 4 we report two examples of comparison using the two methods, and in figure 5 we show the results obtained using all the three methods we have (Direct Comparison, through Predefined Set Comparison, and Modification Comparison) applied to the same two molecules.

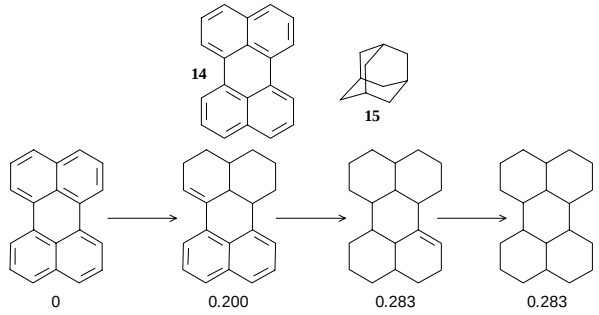

Figure 4. Molecule comparison through their modification.

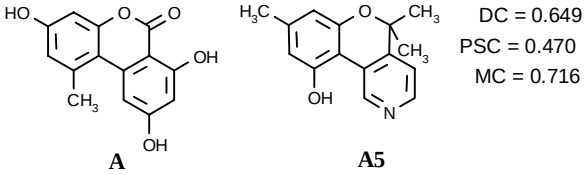

Figure 5. Molecule comparison using three methods.

The examples show comparison of molecules of different kinds, but it must be clear that we can compare any molecule pair, even very different, with our systems.

The Application of Similarity to Complex Problems

The development of similarity measuring methods is very useful for structure comparison; but, more interesting, is their use in complex fields, where the number of unpredictable variables is large.

Our previous experience in the area of computer aided organic synthesis planning has stimulated our interest towards the possibilities offered by the application of similarity. It is evident that the use of analogy is very common in the planning activity, consequently we can easily imagine the introduction of the similarity tool. Though the planning of synthesis is a complex operation that includes multilevel choices accounting for the number and the goodness of the results, we developed methods of similarity application to the comparison of alternative synthetic paths both inside a single synthesis plan (Sello, G., and Termini, M. 1997) and between different plans. (Sello, G. 1998b) This required a great effort in the organisation of the comparisons. We also introduced methods for the evaluation of compound reactivity based on their similarity together with the corresponding similarity measures. For example, we realise the comparison of the two syntheses reported in figure 6. Despite the clear diversity of these

molecules it is still possible to have some hints of the similarity of their synthetic routes. Obviously, the result cannot be expressed using only one number, because the synthetic important aspects are many and have different meaning; e.g. structure simplification is a strategic aspect, whereas reaction efficiency is a tactic aspect.

The principal hint we gathered from our attempts of modelling complex problems concerns the inadequacy of the common methodologies of data analysis. When the problem is complex and the means of its description are scarce, when we have to cohabit with the uncertainty of the experimental measures and with the approximations of the calculated measures, the resort to more fertile and innovative methodologies of data analyses is sure. Here, the characteristics of expert systems, or more generally of artificial intelligence, can be of great help.

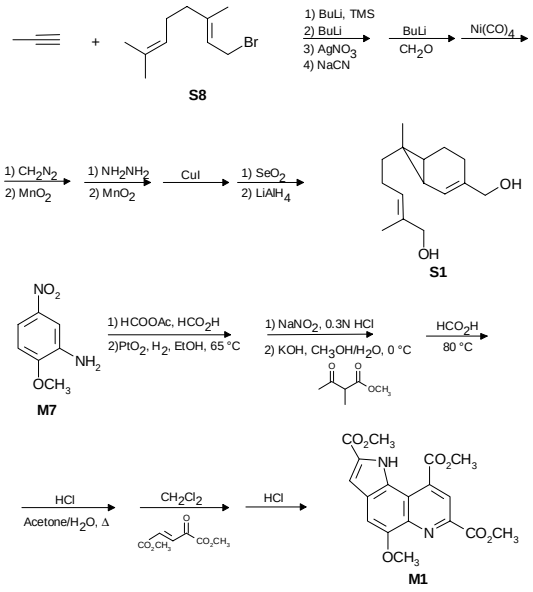

Figure 6. Syntheses comparison using similarity.

The Application of Similarity to Compound Toxicity Prediction

There remains to add some considerations on the possibility of application of similarity to the theoretical evaluation of compound toxicity. If we consider toxicity indistinguishable from any other biological activity the possibilities offered by similarity are self evident. In fact, structure comparison is a common method to assess the activity of unknown from that of known compounds. Because comparison by similarity is a good methodology it is highly probable that it can be used also in this field. However, between the many possibilities it is better to select those that permit the location of the substructures responsible of the activity in addition to a general biological behaviour. In this way, it could be even possible to track the activities related to different biological mode of action.

More, the case of toxicity is evidently a complex problem, where there is the need of considering many different and concurrent aspects of the response of the organism to a substance; thus, the use of many different similarities can be envisaged as a powerful method of prediction. But here the tools used to analyse and validate the results must be carefully chosen because the final answer must be as reliable as possible. Most of the AI approaches can be suggested to care of this vital point of the theoretical prediction of toxicity; we are looking forward to experimenting this possibility.

Acknowledgements. I would like to thank all the present and past collaborators. Partial financial support by the Consiglio Nazionale delle Ricerche and by the Ministero dell'Universita' e della Ricerca Scientifica e Tecnologica, is gratefully acknowledged.

References

Baumer, L., and Sello, G. 1992. A New Method for the Calculation of Natural Bond Polarity Using Molecular Electronic Energy. *J. Chem. Inf. Comput. Sci.* 32:125-130.

Leoni, B., and Sello, G. 1995. A Proposal Toward the Identification of Substructure Electronic Similarity. In *Molecular Similarity and Reactivity from Quantum Chemical to Phenomenological Approaches*, 267-289. Carbo', R. ed. Dordrecht: Kluwer Academic Publ.

Polloth, C., and Mangelsdorf, I. 1997. Commentary on the Application of (Q)SAR to the Toxicological Evaluation of Existing Chemicals. *Chemosphere* 35:2525-2542.

Sello, G. 1992. A New Definition of Functional Groups and a General Procedure for Their Identification in Organic Structures. *J. Am. Chem. Soc.* 114:3306-3311.

Sello, G. 1998a. Similarity Measures: Is It Possible to Compare Dissimilar Structures? *J. Chem. Inf. Comput. Sci.* 38:691-701.

Sello, G. 1998b. Similarity in Organic Synthesis Design: Comparing the Syntheses of Different Compounds. In *Advances in Molecular Similarity*, 2:135-149. Carbo', R., and Mezey, P. eds. London: JAI Press Inc.

Sello, G., and Termini, M. 1996a. Automatic Search for Substructure Similarity. Canonical versus Maximal Matching. Topological versus Spatial Matching. In *Advances in Molecular Similarity*, 1:213-241. Carbo', R., and Mezey, P. eds. London: JAI Press Inc.

Sello, G., and Termini, M. 1996b. Using a Canonical Matching to Measure the Similarity between Molecules: the Taxol and Combretastatine AB1 Case. In *Advances in Molecular Similarity*, 1:243-266. Carbo', R., and Mezey, P. eds. London: JAI Press Inc.

Sello, G., and Termini, M. 1997. Organic Synthesis Planning: Some Hints from Similarity. *Tetrahedron* 53:3729-3756.