A NOVEL APPROACH TO THE CHARACTERIZATION OF MOLECULAR SIMILARITY. THE 2ND ORDER SIMILARITY INDEX

Robert PONEC and Martin STRNAD

Institute of Chemical Process Fundamentals,
Czechoslovak Academy of Sciences, 165 02 Prague 6-Suchdol

Received June 14, 1989
Accepted August 23, 1989

Dedicated to Professor Otto Exner on the occasion of his 65th birthday.

The topological similarity index r_{AB} introduced recently is generalized by using reduced 2nd order density matrices. The comparison with the results of our previous study demonstrates enhanced ability to discriminate between allowed and forbidden reactions.

In the course of its historical development, chemistry has introduced a number of very useful qualitative terms, notions and concepts, on the basis of which an immense wealth of experimental material could be empirically classified. Among such useful concepts is that of molecular similarity, which gave rise to notions such as isoelectronicity, isostericity, aromaticity, etc., in attempts to reduce the very broad idea of similarity to the similarity of some specific molecular property only (electronic structure, steric arrangement, etc.). Although the importance of these qualitatively defined characteristics is undisputable, the need is increasingly felt of obtaining some detailed characteristics that would add a new, quantitative dimension to the qualitative similarity concept.

An important field where quantitative characteristics of molecular similarity are finding application is the rapidly developing mathematical chemistry and its efforts to set up programs for computer designed syntheses[1,2]. Molecular similarity is also used by manufacturers of pharmaceutical drugs and agrochemical products for rationalization of syntheses of biologically active substances.

Among the most widely used characteristics, employed specifically for this purpose, is the so-called similarity index r_{AB}, introduced by Carbo and coworkers[3] based on the density matrices ϱ_A, ϱ_B. Concurrently with efforts to apply this index in practice[4-7], new approaches to the characterization of molecular similarity have been developing with the aim to eliminate some limitations to the applicability of the initial index r_{AB}. For instance, Cooper and Allan[8] suggest similarity indices derived from density matrices not in the coordinates but in the impulse representation. An alternative approach, which is the subject of the present paper, consists in a generalization of the so-called topological similarity index, introduced previously[7].

The modification suggested does not rest in a change in the representation of the density matrix; instead, we pass from 1st order to 2nd order density matrices. The philosophy of this generalization is based on a simple idea as follows.

In the quantum mechanics concept, all information on the structure of a molecule is contained in its wave function. Part of this information, however, loses on passing from wave functions to density matrices, i.e. on integrating over coordinates of a certain number of electrons. The extent of this information loss depends on the degree of reduction, i.e. the number of coordinates over which we integrate, and is highest when going to 1st order density matrices. Thus, we can justly assume that the information content of the similarity index will be higher if 2nd order density matrices are used, these matrices characterizing the electronic structure more completely than 1st order matrices.

The most important difference resulting from the use of the second order density matrices consists in that they include, even if only partialy, the electron correlation. This inclusion concerns only the so-called Fermi correlation which is the correlation between the electrons with the same spin[9]. Although this type of correlation has no energetic impact (correlation energy) the second order density matrices are affected. Hence, the use of the 2nd order density matrix can be of use in all cases where electronic correlation plays a major role, in the field of chemical reactivity for instance. Our previous work[7] oriented in this direction, in particular aimed at an alternative formulation of the Woodward–Hoffmann rules, revealed some limitations due to which the original similarity index r_{AB} fails to discriminate between allowed and forbidden reactions in some cases. In this paper we are going to demonstrate that these limitations are eliminated by the use of the 2nd order similarity index with its higher information content.

It is proper to mention that the above limitations can be also eliminated in alternative ways, e.g., by using the integral similarity index J (refs[10,11]); still, we consider the introduction of the 2nd order similarity index useful. It has several advantages, for instance, it is more easy to determine than the integral index J, and moreover, also this 2nd order similarity index can be further generalized to obtain the integral characteristics, whereupon its predicting ability is additionally increased.

THEORETICAL

Let us first consider some specific problems associated with the generalization suggested. First, it is clear that the definition of the original index r_{AB} must be modified to respect the inclusion of the 2nd order density matrices. The straightforward generalization is by Eq. (*1*)

$$g_{AB} = \frac{\int \tilde{\varrho}_A \tilde{\varrho}_B \, d\tau_1 \, d\tau_2}{\left(\int \tilde{\varrho}_A^2 \, d\tau_1 \, d\tau_2\right)^{1/2} \left(\int \tilde{\varrho}_B^2 \, d\tau_1 \, d\tau_2\right)^{1/2}}, \qquad (1)$$

where $\tilde{\varrho}_A$ and $\tilde{\varrho}_B$ are the reduced spinless 2nd order density matrices and $d\tau_1$ and $d\tau_2$ are volume elements with respect to the first and second electrons, respectively. Next, since we wish to arrive at a straightforward generalization of the initial formalism[7], the practical calculation of the similarity index g_{AB} must proceed from the same assumptions and approximations as that of the index r_{AB}. The basis for this is again in the topological approach employing the overlap determinant method[12], where the structure of the two molecules concerned, A and B, is described by approximate wave functions ϕ_A and ϕ_B, respectively, constructed in the form of Slater determinant from the molecular orbitals ϕ_i, ψ_j as follows:

$$\phi_A = |\phi_1\bar{\phi}_1\phi_2\bar{\phi}_2 \ldots \phi_n\bar{\phi}_n| \tag{2a}$$

$$\phi_B = |\psi_1\bar{\psi}_1\psi_2\bar{\psi}_2 \ldots \psi_n\bar{\psi}_n| \,. \tag{2b}$$

These orbitals then are described, as in the conventional LCAO approximation, by means of linear combinations of atomic orbitals:

$$\phi_i = \sum_{\mu} a_{\mu i}\chi_\mu \tag{3a}$$

$$\psi_j = \sum_{v} b_{vj}\chi'_v \,, \tag{3b}$$

where the prime in Eq. (*3b*) is used to show explicitly that the (real) basis χ' in which orbitals of molecule B are expressed may be generally different from the (real) basis χ describing molecule A.

Within the framework of such a model it can be demonstrated that the 2nd order spinless density matrices ϱ are described by the equations

$$\tilde{\varrho}_A(1, 2) = \sum_{\alpha\beta}\sum_{\mu\sigma} q^{A}_{\alpha\beta\mu\sigma}\chi_\alpha(1)\,\chi_\beta(1)\,\chi_\mu(2)\,\chi_\sigma(2) \tag{4a}$$

$$\tilde{\varrho}_B(1, 2) = \sum_{\alpha\beta}\sum_{\mu\sigma} q^{B}_{\alpha\beta\mu\sigma}\chi'_\alpha(1)\,\chi'_\beta(1)\,\chi'_\mu(2)\,\chi'_\sigma(2) \,, \tag{4b}$$

and since we have

$$\tilde{\varrho}_X(1, 2) = \varrho_X(1)\,\varrho_X(2) - \varrho^2_X(1, 2)/2 \quad (X = A, B) \,, \tag{5}$$

where

$$\varrho_A(1) = 2\sum_{i}^{\text{occup}} \phi^2_i(1) \tag{6a}$$

$$\varrho_B(1) = 2\sum_{j}^{\text{occup}} \psi^2_j(1) \tag{6b}$$

$$\varrho_A(1, 2) = 2\sum_{i}^{\text{occup}} \phi_i(1)\,\phi_i(2) \tag{6c}$$

$$\varrho_B(1,2) = 2 \sum_j^{\text{occup}} \psi_j(1)\,\psi_j(2)\,, \tag{6d}$$

the expressions for the matrix elements $q^A_{\alpha\beta\mu\sigma}$ and $q^B_{\alpha\beta\mu\sigma}$, viz.

$$q^A_{\alpha\beta\mu\sigma} = p^A_{\alpha\beta}p^A_{\mu\sigma} - (p^A_{\alpha\mu}p^A_{\beta\sigma} + p^A_{\alpha\sigma}p^A_{\beta\mu})/4 \tag{7a}$$

$$q^B_{\alpha\beta\mu\sigma} = p^B_{\alpha\beta}p^B_{\mu\sigma} - (p^B_{\alpha\mu}p^B_{\beta\sigma} + p^B_{\alpha\sigma}p^B_{\beta\mu})/4 \tag{7b}$$

can be written in terms of matrix elements $p^A_{\alpha\beta}$, $p^B_{\alpha\beta}$ of the 1st order density matrix as

$$p^A_{\alpha\beta} = 2 \sum_i^{\text{occup}} a_{\alpha i}a_{\beta i} \tag{8a}$$

$$p^B_{\alpha\beta} = 2 \sum_j^{\text{occup}} b_{\alpha j}b_{\beta j}\,. \tag{8b}$$

Although these relations offer a general methodological basis for the calculation of the index g_{AB}, practical application of the formalism suggested again, as was the case with the index r_{AB}, encounters difficulties associated with problems of invariance with respect to the mutual position and orientation of molecules A, B. Similarly as before[7], it turns out that these problems can be eliminated in a simple manner within the topological approximation based on the overlap determinant method. This elimition is based on the use of the so-called assignment tables[12], describing the mutual relation of the AO bases χ, χ' in Eqs (*4a*), (*4b*). Actually, the mutual relation of the two bases can be conveniently expressed[7] in the alternative matrix form

$$\chi'_\alpha = \sum_\beta \tau_{\alpha\beta}\chi_\beta \tag{9}$$

which can be also employed in the case of the index g_{AB}. Really, by means of Eq. (*9*), Eq. (*4b*) can be modified to the form

$$\tilde{\varrho}_B(1,2) = \sum_{\alpha\beta}\sum_{\mu\sigma} \bar{q}^B_{\alpha\beta\mu\sigma}\chi_\alpha(1)\,\chi_\beta(1)\,\chi_\mu(2)\,\chi_\sigma(2)\,, \tag{10}$$

where, analogously to Eq. (*7b*), we have

$$\bar{q}^B_{\alpha\beta\mu\sigma} = \bar{p}^B_{\alpha\beta}\bar{p}^B_{\mu\sigma} - (\bar{p}^B_{\alpha\mu}\bar{p}^B_{\beta\sigma} + \bar{p}^B_{\alpha\sigma}\bar{p}^B_{\beta\mu})/4 \tag{11}$$

and, similarly as in ref.[7],

$$\bar{p}^B_{\alpha\beta} = \sum_{\mu\sigma} \tau^{-1}_{\alpha\mu}\,p^B_{\mu\sigma}\tau_{\sigma\beta}\,. \tag{12}$$

Using Eqs (*7a*), (*10*) and the same ZDO-like approximation as previously[7], we can transform Eq. (*1*) to the form

$$g_{AB} = \frac{\sum_{\mu\sigma}\sum_{\alpha\beta} q^{A}_{\alpha\beta\mu\sigma}\bar{q}^{B}_{\alpha\beta\mu\sigma}}{[\sum_{\alpha\beta}\sum_{\mu\sigma}(q^{A}_{\alpha\beta\mu\sigma})^2]^{1/2}\,[\sum_{\alpha\beta}\sum_{\mu\sigma}(\bar{q}^{B}_{\alpha\beta\mu\sigma})^2]^{1/2}} \qquad (13)$$

which can be further rearranged to the final form (*16*) by using Eqs (*7a*), (*11*), the idempotency relation (*14*) and identity (*15*). ($\boldsymbol{P}_A, \bar{\boldsymbol{P}}_B$ symbolize the reduced 1st order density matrices, N is number of the electrons in the system):

$$\sum_{\mu} p_{\alpha\mu}p_{\beta\mu} = 2p_{\alpha\beta}\,, \qquad (14)$$

$$\sum_{\alpha\beta} p^2_{\alpha\beta} = 2N\,, \qquad (15)$$

$$g_{AB} = \frac{9\,\mathrm{Tr}^2(\boldsymbol{P}_A\bar{\boldsymbol{P}}_B) - 7\,\mathrm{Tr}\,(\boldsymbol{P}_A\bar{\boldsymbol{P}}_B\boldsymbol{P}_A\bar{\boldsymbol{P}}_B)}{4N(9N - 14)}\,. \qquad (16)$$

This relation, which, naturally, only holds true if the wave functions are described by single determinants, is very useful, demonstrating that the 1st order density matrices are sufficient also for the calculation of the 2nd order similarity index; this is very convenient from the numerical point of view. Owing to its simple calcula-

TABLE I

Calculated values of the 1st and 2nd order similarity indices for some allowed and forbidden pericyclic reactions

Reaction	Mechanism		r_{AB}	g_{AB}
2 + 2 addition	s, a	allowed	0·500	0·250
	s, s	forbidden	0·500	0·091
Diels–Alder reaction	s, s	allowed	0·574	0·298
	s, a	forbidden	0·574	0·272
Butadiene/cyclobutene	con	allowed	0·723	0·524
	dis	forbidden	0·500	0·091
Hexatriene/cyclohexadiene	dis	allowed	0·759	0·566
	con	forbidden	0·658	0·358
Octatetraene/cyclooctatriene	con	allowed	0·718	0·503
	dis	forbidden	0·661	0·401

tion, the similarity index can find application in all cases where increased sensitivity demands with respect to differences in the electronic structure can be relevant, such as our case of pericyclic reactivity. In the following discussion we are going to demonstrate that the expected increase in the information content associated with an increase in the discriminating ability of the 2nd order similarity index really takes place.

DISCUSSION

For the comparison we shall use basically the same series as has served in our previous work[7] for testing and discussing the similarity index r_{AB}. This allows us to omit specification of the technical details of the calculations, which can be found in ref.[7], and concentrate on the discussion of the calculated g_{AB} values and their comparison with the original r_{AB} values.

The data for a series of electrocyclic and cycloaddition reactions are given in Table I. We see that generally the g_{AB} values are lower than the r_{AB} values. This is a consequence of the higher information content of the former; really, it is clear that the more aspects of the electronic structure of molecules we are able to describe, the lower is the probability that all of the aspects for two different molecules are so close to each other that a sufficiently high similarity emerges. The fact that the g_{AB} values are lower than the r_{AB} values is not, therefore, very significant on its own. What interests us more is the mutual comparison of the g_{AB} values for allowed and forbidden reactions. This comparison gives evidence that, in agreement with the principle of least motion, the extent of electron reorganization accompanying the R → P transformation is lower in allowed reactions than in forbidden ones. In this context it is also of importance that, owing to the increase in its information content arising from the use of the 2nd order density matrices, the g_{AB} index really makes it possible to discriminate between allowed and forbidden reactions even in the case of cycloadditions, in which the r_{AB} index fails. The fact that most of the additional information contained in the g_{AB} index is related to electronic correlation suggests that this index may find application not only for an alternative formulation of the Woodward–Hoffmann rules as indicated, but also as a universal tool for studying effects associated with the mutual coupling of electronic motions in the field of chemical reactivity in general.

REFERENCES

1. Ugi J., Bauer J., Brandt J., Friedrich J., Gasteiger J., Jochum C., Schubert W.: Angew. Chem., Int. Ed. *18*, 111 (1979).
2. Koča J., Kratochvíl M., Kvasnička V., Matyska L., Pospíchal J. in: *Synthon Model of Organic Chemistry and Synthesis Design. Lecture Notes in Chemistry*, Vol. 51. Springer, Berlin 1989.
3. Carbo R., Leyla L., Arnau M.: Int. J. Quantum Chem. *17*, 1185 (1980).

4. Hodgkin E. E., Richards W. G.: J. Chem. Soc., Chem. Commun. *1986*, 1342.
5. Bowen-Jenkins P. E., Cooper D. L., Richards W. G.: J. Phys. Chem. *89*, 2195 (1985).
6. Hodgkin E. E., Richards W. G.: Int. J. Quantum Chem., Quantum Biol. Symp. *14*, 105 (1987).
7. Ponec R.: Collect. Czech.`Chem. Commun. *52*, 555 (1987).
8. Cooper D. L., Allan N. L.: J. Computer Aided Design, in press.
9. Kutzeknigg W.: Top. Curr. Chem. *41*, 31 (1973).
10. Ponec R.: Z. Phys. Chem. (Leipzig) *168*, 1180 (1987).
11. Ponec R.: Z. Phys. Chem. (Leipzig) *270*, 365 (1989).
12. Ponec R.: Collect. Czech. Chem. Commun. *49*, 455 (1984).

Translated by P. Adámek.