CROATICA CHEMICA ACTA CCACAA **66** (1) 123–127 (1993)

ISSN 0011–1643
CCA–2123

Conference Paper

Position Invariant Index for Assessment of Molecular Similarity

Robert Ponec and Martin Strnad

Institute of Chemical Process Fundamentals, Czechoslovak Academy of Sciences, Prague 6, Suchdol 2, 165 02, Czechoslovakia

Received August 17, 1992

In this study the idea of the topological similarity index has been generalized so that its applicability reaches beyond a simple HMO model to semiempirical or even *ab initio* methods. The greatest advantage of this new index is its complete positional invariance, which makes it possible to avoid time consuming optimization with respect to the mutual position of molecules. As a consequence of this invariance, the index can find a number of interesting applications, especially in the design of compounds with desired biological properties. The applicability of the index has been numerically tested on a series of isosteric and isoelectronic molecules.

INTRODUCTION

Creation of theoretical models describing the behaviour of the studied systems is a general process that accompanies the increasing exploitation of various mathematical methods and procedures in all areas of natural sciences. In contemporary chemistry, this general tendency finds its reflection in continuing attempts at specifying the exact meaning of various useful, but originally rather intuitively defined, notions and concepts. An extremely useful such concept is that of molecular similarity. Although the applicability of this concept is extremely broad and heterogeneous,[1] the great interest devoted in recent years to the exploitation of molecular similarity undoubtedly has its roots in the increasing, practically motivated effort of pharmaceutical and agrochemical companies to rationalize the search for compounds with desired biological properties. Exploitation of the so-called similarity index is an important aspect of this effort.[2–10]

However, in spite of this extensive use, these indices have one important conceptual disadvantage: their values depend on the distance and the mutual orientation of the molecules. In order to remedy this major disadvantage, several different procedures have been proposed. Most of them are based on optimizing the mutual position of the molecules so as to maximize the value of the similarity index. However, such a straight-

forward procedure is quite time consuming and, moreover, it usually gives no guarantee that the found optimum does indeed correspond to the global and not only to the local maximum.

For that reason, some other techniques have also emerged, which attempt to use other means to solve the problem of the positional dependence of the similarity index. As an example, we may mention the study by Cooper and Allan,[8] in which considerable reduction of positional dependence results from the use of momentum, rather than the usual coordinate representation of density matrices. Another example of the positional invariant similarity index is the so-called topological similarity index[11] the introduction of which is based on incorporating Carbo's[3] original index into the framework of the overlap determinant method.[12]

Although it has found a number of interesting applications, especially in the field of pericyclic reactivity, it is nevertheless a disadvantage that this index was originally introduced only at the level of a simple HMO model, where only one atomic orbital of the p-type is localized on each atom. Such a model is, of course, only very crude and its applicability is restricted only to the simplest model calculations. Our aim in this study is to overcome this limitation and to generalize the original, from the HMO model derived index, so as to make it also applicable to more sophisticated semiempirical or even *ab initio* methods. The result of such a generalization would be an index combining the generality of Carbo's definition with strict positional invariance, so that time consuming positional optimization can be completely avoided. In addition to this saving of computer time, positional invariance is also a very convenient feature for subsequent applications of the index in the design of biologically active compounds.

THEORETICAL

Since the proposed generalization is directly related to the so-called topological similarity index introduced in our previous study of some time ago,[11] we consider it convenient to make a brief recapitulation of the basic ideas from which the above index is derived.

The basis of this index is the incorporation of Carbo's[3] original definition (Eq. 1) into the framework of the so-called overlap determinant method.[12]

$$r_{AB} = \frac{\int \rho_A \rho_B \, d\upsilon}{\left(\int \rho_A^2 \, d\upsilon\right)^{1/2} \left(\int \rho_B^2 \, d\upsilon\right)^{1/2}} \tag{1}$$

Within the framework of this method, the basic problem of the determination of molecular integrals $I_{\mu\nu\lambda\sigma}$ (2), which are the source of the non-invariance of the similarity index, has been solved by the unitary transformation matrix $\boldsymbol{T}$ (Eq. 3), describing the interrelation of AO basis sets $\{\chi^A\}$, $\{\chi^B\}$ in both individual molecules.

$$I_{\mu\nu\lambda\sigma} = \int \chi_\mu^A \chi_\nu^A \chi_\lambda^B \chi_\sigma^B \, d\tau \tag{2}$$

$$\chi_\lambda^B = \sum_\sigma \boldsymbol{T}_{\lambda\sigma} \chi_\sigma^A \tag{3}$$

On the basis of this matrix, together with a subsequent ZDO-like approximation (4), the original definition equation (1) can be

$$I_{\mu\nu\lambda\sigma} = \delta_{\mu\lambda}\,\delta_{\nu\sigma} \tag{4}$$

rewritten in the final form (5),

$$r_{AB} = \frac{\mathrm{Tr}\,\boldsymbol{P}_A\,\overline{\boldsymbol{P}}_B}{2N} \tag{5}$$

where the prime over matrix $\overline{\boldsymbol{P}}_B$ denotes the unitary transformation (6) with matrix $\boldsymbol{T}$.

$$\overline{\boldsymbol{P}}_B = \boldsymbol{T}^{-1}\,\boldsymbol{P}_B\,\boldsymbol{T}\,. \tag{6}$$

While in the case of the original topological index,[9] defined at the level of the HMO method, matrix $\boldsymbol{T}$ is very simple and could be written down without any calculations, in the case of more sophisticated MO methods the form of matrix $\boldsymbol{T}$ is more complex and has to be determined numerically. The criteria for the determination of this matrix can, of course, be formulated rather arbitrarily, but out of the various possible choices the most natural is probably the one requiring matrix $\boldsymbol{T}$ to be determined from the condition of the maximization of index r_{AB}.

It is interesting to note that the same problem, albeit with a slightly different motivation, was analyzed some time ago by Trindle[13] and even if an algorithm allowing the determination of matrix $\boldsymbol{T}$ according to the above criterion was proposed by him, the close relation of his procedure to the determination of the similarity index has so far remained unnoticed.

Our aim in this study is to exploit Trindle's algorithm based on the iterative optimization of matrix $\boldsymbol{T}$ for determining the positionally invariant similarity index. The above procedure implemented into the interactive computer program which uses the MNDO calculated density matrices as input will be numerically demonstrated using the example of similarity indices of several selected molecules. The greatest advantage of the proposed approach is the complete positional invariance of the similarity index, which thus becomes a really significant and objective measure of the resemblance of molecular structures. In addition to this important feature another great advantage of the above index lies in the ease of its calculation, which is incomparable with time consuming methods requiring positional optimization. In order to illustrate the speed of calculations, it is possible to specify that, for example for medium sized molecules with let us say 10 heavy atoms, the calculation requires roughly 90 s on a PC of the AT series with a mathematical coprocessor.

RESULTS AND DISCUSSION

The above procedure was numerically tested on a series of selected molecules involving the isosteric set $CH_3CH_2CH_3$, CH_3SCH_3, CH_3OCH_3 and the isoelectronic series CH_3CH_3, CH_3NH_2, CH_3OH, CH_3F, CH_3Cl. Density matrices representing the input for Trindle's algorithm were generated by the standard MNDO method[14,15] for fully optimized molecular geometries. The calculated values of similarity indices for both series of compounds are summarized in Tables I and II.

TABLE I

Calculated values of similarity index $\mathbf{r}_{AB}$ *for a series of isoelectronic molecules* CH_3XH_n *(X = C, N, O, F, Cl;* n = *0—3).*

CH_3CH_3	CH_3NH_2	CH_3OH	CH_3F	CH_3Cl
1.0000	0.9256	0.8518	0.7802	0.7784
	1.0000	0.9346	0.8691	0.8670
		1.0000	0.9404	0.9383
			1.0000	0.9947
				1.0000

TABLE II

Calculated values of similarity index $\mathbf{r}_{AB}$ *for the series of isosteric molecules* $CH_3CH_2CH_3$, CH_3SCH_3 *and* CH_3OCH_3.

$CH_3CH_2CH_3$	CH_3SCH_3	CH_3OCH_3
1.0000	0.8972	0.8960
	1.0000	0.9965
		1.0000

Let us now discuss the values from these Tables. The first interesting result concerns systematic trends in the similarity indices for the series CH_3XH_n (X = C, N, O, F, Cl; n = 0 – 3), characterized by a regular decrease in the number of hydrogens, in the XH_n group. In agreement with intuitive expectations, this regular decrease finds its reflection in a regular decrease in similarity indices for the pairs CH_3CH_3/CH_3XH_n, CH_3NH_2/CH_3XH_2, CH_3OH/CH_3XH_n and $CH_3F(Cl)/(CH_3XH_n$. The greatest similarity is then observed for the pair CH_3F/CH_3Cl, which differs only in the substitution of Cl for F. Further interesting conclusions can be deduced from the comparison of a series of isosteric molecules $CH_3CH_2CH_3/CH_3SCH_3$ and $CH_3CH_2CH_3/CH_3OCH_3$. As it can be seen from Table II, the greatest similarity is observed for the pair propane/dimethylsulphide, whereas for the analogous pair propane/dimethilether the value of $\boldsymbol{r}_{AB}$ is slightly lower. This result, which was also reported in the study,[8] seems entirely reasonable since it reflects the well known experience of empirical drug design that the replacement of $-CH_2-$ by –S– usually leads to similar biological properties, whereas the same is frequently not true for the pair $-CH_2-$ /–O–. Summarizing the above results, we would like to conclude by expressing our belief that the proposed generalized index can become a simple useful means for evaluating molecular similarity, and that its future systematic use will contribute to a broader exploitation of the ideas of similarity in various fields of both theoretical and practically oriented chemical research.

Acknowledgement. — The authors thank Mr. Robin Healy, the British Council representative in Prague, for reading the manuscript and for linguistic corrections.

REFERENCES

1. D. Rouvray, in: *Concepts and Applications of Molecular Similarity*, M. A. Johnson and G. M. Maggiora, Wiley, Eds., New York 1990. Chap. 2, p. 15.
2. O. E. Polansky and G. Derflinger, *Int. J. Quant. Chem.* **1** (1967) 379.

3. R. Carbo, L. Leyda, and M. Arnau, *Int. J. Quant. Chem.* **17** (1980) 1185.
4. P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.* **89** (1985) 2195.
5. E. E. Hodgkin and W. G. Richards, *J. Chem. Soc., Chem. Commun.* (1985) 1342.
6. R. Carbo and B. Calabuig, *Comp. Phys. Commun.* **14** (1987) 105.
7. E. E. Hodgkin and W. G. Richards, *Int. J. Quant. Chem., Quant. Chem. Symp.* **14** (1987) 105.
8. D. L. Cooper and N. Allan, *J. Comp. Aided Mol Design* **3** (1989) 253.
9. C. Amovilli and R. McWeeny, *J. Mol. Struct. (Theochem)* **227** (1989) 1.
10. J. Cioslowski and E. D. Fleischmann, *J. Amer. Chem. Soc.* **112** (1991) 64.
11. R. Ponec, *Coll. Czech. Chem. Commun.* **52** (1987) 555.
12. R. Ponec, *ibid.* **49** (1984) 455.
13. C. Trindle, *J. Amer. Chem. Soc.* **92** (1971) 3251.
14. M. J. S. Dewar and W. Thiel, *J. Amer. Chem. Soc.* **99** (1977) 4899.
15. M. J. S. Dewar and W. Thiel, *ibid.* **99** (1977) 4907.

SAŽETAK

Procjena molekulske sličnosti s pomoću prostorno invarijantnog indeksa

Robert Ponec i Martin Strnad

Ideja topologijskih indeksa sličnosti poopćena je tako da oni budu primjenljivi u okviru semiempirijskih pa čak i *ab initio* računa. Najveća prednost novog indeksa jest njegova potpuna prostorna invarijantnost što omogućuje niz zanimljivih primjena, posebno u projektiranju molekula određenih bioloških svojstava. Primenljivost indeksa numerički je iskušana na nizu izosteričkih i izoelektronskih molekula.