communications in mathematical
match
and in computer chemistry

ISSN 0340-6253

no. 35, March 1997

MATCDY (35) 29-40 (1997)

The Molecular Transform as a Tool for Quantification of Molecular Similarity

Gábor NÁRAY-SZABÓ and Veronika HARMAT

Department of Theoretical Chemistry, Loránd Eötvös University Budapest, P.O. Box 32, H-1518 Budapest 112, Hungary, E-mail: ETHE003@URSUS.BKE.HU

(Received: July 1996)

Abstract. We call the attention to the experimentally measurable molecular transform as a basis for the quantitative characterisation of molecular similarity. Beside the full mathematical expression simplified topological forms can be derived allowing to calculate various metric distances between molecular transforms very rapidly. Such distances may be used for defining various groups of molecules possessing similar physical, chemical or biological properties. A further application is isostructurality of crystals where unit cells or asymmetric units can be compared quantitatively in order to derive general rules of crystal packing.

1. Introduction

The concept of molecular similarity gains more and more interest in theoretical chemistry, especially chemical information theory and quantitative structure-property studies [1,2]. Beside qualitative considerations it is also possible to define various quantitative measures of molecular similarity, here we mention only a few of the numerous examples available in the literature [3-9]. While it is important to find an adequate and quantitative mathematical description, there is also a need for formulae that can be evaluated rapidly by

the computer thus allowing quick comparisons in large molecular libraries. In the present paper we give an overview on a definition and its applications that is based on molecular transforms and may fulfil the above goals.

The molecular transform is obtainable experimentally, e.g. by diffraction methods [10], as well as theoretically if the three-dimensional atomic co-ordinates are available. Its application to rational drug design has been first proposed by Soltzberg and Wilkins [3] and somewhat later we proposed a topological form allowing the rapid calculation of abstract molecular distances that are used for measuring similarity [5]. Recently King and co-workers addressed the problem and defined 2D and 3D molecular transforms which they used for the derivation of various structure-property relationships [9].

In the following we deal with the application of various simplified forms of the molecular transform to the definition of abstract distances between molecules. Theoretical details are given in a recent paper [11] therefore we concentrate on applications to practical problems, like linear regression equations for physical properties of aliphatic alcohols and structure-pharmacological activity studies for various classes of molecules. Another application is the study of isostructurality of digitoxigenin analogues [12] that allows to deduce some relationship between molecular structure and crystal packing.

2. Theory

The molecular transform is written as a Fourier transform of the scattered radiation observed in an electron diffraction experiment [10]

$$I(s) = K \sum_{i<j} f_i f_j \int P_{ij}(r)\sin(sr)/sr\,dr \qquad (1)$$

where K is a constant, f_i and f_j are the form factors, $P_{ij}(r)$ is the probability distribution of the vibrational variation in the distance between atoms i and j. $s = 4\pi\lambda\sin(\theta/2)$ where λ is the wavelength of the electron beam and θ is the scattering angle. With $K = 1$ and $P_{ij}(r) = \delta(r - r_{ij})$ we get a simplified expression

$$I(s) = K \sum_{i<j} f_i f_j F(x) \tag{2}$$

with $x = sr_{ij}$ and $F(x) = \sin(x)/x$. We may define r_{ij} both as the geometric distance and the shortest path in the molecular graph between atoms i and j in order to obtain a 3D or a 2D representation of the molecule, respectively. While in the original definition f_i is equal to the atomic number, it is possible to use other atomic parameters, e.g. the net charge to obtain generalisations of Eq. (2).

In order to define abstract distances between molecules a and b we use the following formula [11]

$$R^2_{ab} = 2(1 - N^2_{ab}/N_{aa}N_{bb}) \tag{3}$$

with

$$N^2_{ab} = \int I_a(s) I_b(s) \mathrm{d}s \tag{4}$$

Using the definition of the molecular transform in Eq. (2) N^2_{ab} can be expressed in a closed form

$$N^2_{ab} = \sum^a_{i<j} \sum^b_{k<l} f^a_i f^a_j f^b_k f^b_l \, g_{ab}(D^a_{ij}, D^b_{kl}) \tag{5}$$

where

$$g_{ab}(D^a_{ij}, D^b_{kl}) = \tfrac{1}{2}\pi[\max(D^a_{ij}, D^b_{kl})]^{-1} \tag{6}$$

In the following we call two molecules similar if their abstract distance, as defined in eqs. (3-6) is small. If two molecules are identical, their distance is exactly zero. Since eqs. (5) and (6) are closed mathematical expressions the abstract distance can be calculated quite rapidly.

For the study of the feasibility of molecular transforms for the prediction of various properties we may follow two ways. It is possible to use the abstract distance of a molecule a from a lead, R_{0a}, as a descriptor and derive a linear regression equation correlating R_{0a} with the absolute value of the difference between a certain property measure of molecule a and the lead, $\Delta P_a = |P_0 - P_a|$. Alternatively, we can use R_{0a} to predict whether ΔP_a exceeds a limit or not. In this approach we call successful those predictions for which

$$\Delta P_a > 0.5(\Delta P_{\max} + \Delta P_{\min}) \qquad \text{if} \qquad R_{0a} > 0.5(R_{0\max} + R_{0\min}) \tag{7}$$

or

$$\Delta P_a < 0.5(\Delta P_{\max} + \Delta P_{\min}) \qquad \text{if} \qquad R_{0a} < 0.5(R_{0\max} + R_{0\min}) \tag{8}$$

In eqs. (7) and (8) "min" and "max" refer to minimal and maximal values inside a given group of molecules. This means that if the abstract distance is smaller than the mean between the two extremes, ΔP_a should also be smaller than the mean value of the given property within the group. We call false predictions that do not obey the rule given in eqs. (7) and (8), for a random distribution the ratio of false predictions is just 50 per cent.

3. Quantitative Structure-Property Relationships

The basic premise behind the use of molecular transforms in quantitative structure-property relationships is that the smaller is R_{ab} of eq. (3) for a pair of molecules the closer are their properties to each other. This is exactly true in case of an identity ($R_{ab} = 0$) but does not hold necessarily for larger values of R_{ab} and for all molecular properties. For example, the molecular weight of two structural isomers is exactly the same, though their abstract distance may be quite large. Some molecules may interact with enzymes only via a small fragment that might be very similar in a series of otherwise quite different species. It is therefore important to find those cases where, in fact, smaller R_{ab} indicates larger similarity of a certain molecular property. In the following we present some examples.

3.1. Physical constants

In an earlier publication [11] we investigated the dependence of boiling and melting points, as well as the density and the refraction index of aliphatic alcohols on R_{ab}. Considering methanol as the lead compound (a = 0) we derived the following linear regression equations

$$\Delta BP_b = -1092R_{0b} + 1063 \qquad r = -0.9473 \quad n = 29 \qquad (9)$$

$$\Delta MP_b = -625.1R_{0b} + 610.7 \qquad r = -0.8446 \quad n = 16 \qquad (10)$$

$$\Delta D_b = -0.2373R_{0b} + 0.2397 \qquad r = -0.8230 \quad n = 29 \qquad (11)$$

$$\Delta N_b^3 = -0.00839R_{0b} + 0.00821 \qquad r = -0.9473 \quad n = 29 \qquad (12)$$

where BP, MP, D and N stand for boiling point, melting point, density and refraction index, respectively. Note that in eq. (12) a cubic, instead linear, relationship is given for the dependence of the refractive index on the abstract distance from the lead. It is quite probable that for a number of properties the linear dependence on R_{0b} is not valid therefore more complicated functions should be found. We could not derive similar relationships for substituted phenols like in eqs. (9-12), however, the number of false predictions using eqs. (7-8) was 0, 32 and 43 per cent for BP (20), MP (22) and N (14), respectively (the number of molecules considered is in parentheses). Once again an evidence is provided for the old experience: aliphatic alcohols are most suitable targets for establishing simple mathematical structure-property relationships (see e.g. ref. 13).

3.2. Biological activity

Our main goal with the definition of a new similarity index was to find quantitative structure-biological activity relationships [5]. Using the topological form of the molecular transform and a somewhat different definition of the abstract molecular distance, we tested

the fungicide activity of a series of substituted acetylenic sulphone derivatives where we did not find any correlation. However, for a number of other cases, like the effect of arylamines on *Mycobacterium tuberculosis* (**I**), haemolytic activity (**II**), acute toxicity (**III**) and antibacterial activity (**IV**) of substituted phenols and monoamino-oxidase inhibitory effect of *N*-substituted *N*-methyl-*N*-propargylammonium hydrochloride derivatives (**V**), a fair classification of active and inactive compounds was possible. We considered the most active compound as the lead and compared its activity to that of the others using eqs. (7) and (8). The number of false predictions is the following: **I**: 6 of 18 (33 %), **II**: 5 of 25 (20 %), **III**: 6 of 25 (24 %), **IV**: 4 of 25 (16 %), **V**: 5 of 19 (26 %). Considering the diversity of biological activities studied the results obtained using a single molecular descriptor seem to be acceptable.

The inhibitory power of substituted benzamidine inhibitors of trypsin has also been studied [14]. On the basis of the molecular electrostatic potential patterns around the substituents we divided them into two groups. Group I (3-Me, 3-OH, 3-OMe, 3-OEt, 4-Me, 4-NH_2, 4-OH, 4-OMe, 4-OEt) is characterised by a positive potential pattern all around the substituent, while the potential for Group II (3-NO_2, 3-COMe, 3-COOMe, 3-COOEt, 3-CONHMe, 4-NO_2, 4-COMe, 4-COOMe, 4-COOEt, 4-CONHMe) is, at least partially, negative in this region. It has been found that the pK_i values for Group I are in all, but two, cases (3-OMe, 4-OEt) larger than 4.2, while $pK_i < 4.2$ for Group II in all, but one, case (3-COMe). Using the topological form of R_{ab} as above [5], we could also distinguish between Groups I and II. We found that $R_{ab} < 0.05$ if both *a* and *b* are in the same group (with 1 false prediction for 21 pairs within Group I and 10 false predictions for 66 pairs within Group II), while $R_{ab} > 0.05$ if *a* and *b* belong to different groups (with 15 false predictions for 84 pairs). We checked the validity of the more rigorous eqs. (7) and (8) and found 57 (37 %) false predictions for the union of Groups I and II. This number is obtained as a combination of 24 false predictions out of 45 cases (53%) for eq. (7) and 33 false predictions out of 108 cases (31%) for eq. (8). In other words, in this specific case prediction of pharmacological activities that lie closer to that of the lead is safer than prediction of those lying far from it.

Our last example deals with thiol inhibitors of thermolysin that have been investigated by Bohacek and Martin [16] as potential targets for *de novo* drug design (Figure 1). Plotting the activity differences as compared to the lead (**1**), ΔpK_{ia}, as a function of R_{1a} we obtain the following regression equation

$$\Delta pK_{ia} = 5.68R_{1a} + 1.75 \qquad r = 0.6867 \qquad n = 7 \tag{13}$$

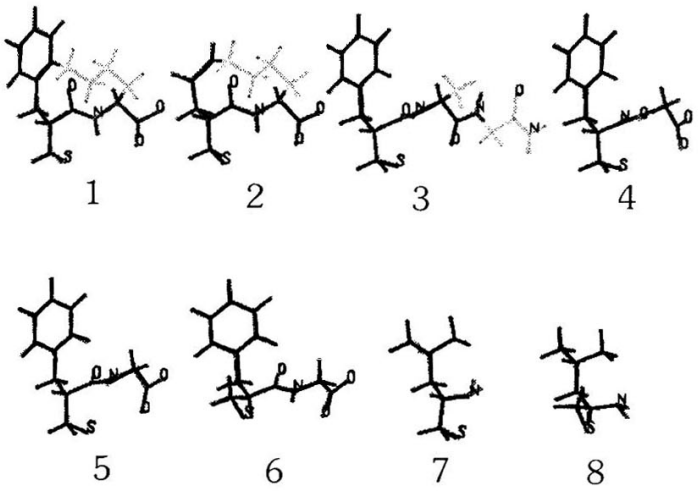

Figure 1. Geometry-optimised structures of some thiol inhibitors of thermolysin, taken from ref. 16 (TRIPOS force field, figures drawn by the SYBYL program [15]). The hypothetical active fragment is drawn by heavy lines. Experimental potencies (pK_{ia}) are as follows. **1:** -1.72, **2:** 0.58, **3:** -0.12, **4:** 0.36, **5:** 0.26, **6:** 0.48, **7:** 0.72, **8:** 1.72.

The correlation can be improved if we consider only the hypothetical active fragment when calculating abstract distances. In this case we obtain the following relationship (cf. Figure 2)

$$\Delta pK_{ia} = 6.91R_{1a} + 1.74 \qquad r = 0.7993 \qquad n = 7 \tag{14}$$

Note that the active fragment is identical for molecules **1**, **5** and **6**, thus the topological (2D)

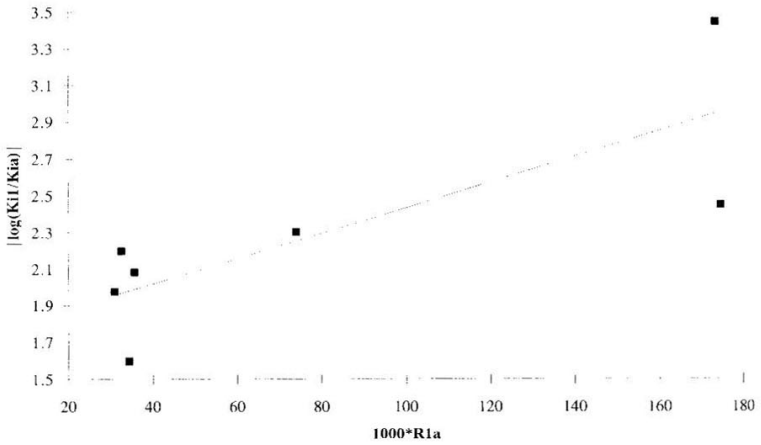

Figure 2. ΔpK_{ia} versus $1000R_{1a}$ plot for thiol inhibitors of thermolysin depicted on Figure 1.

form of the molecular transform yields exactly zero distance between them. However, owing to the absence of the saturated ring in **5** the molecular geometries are different thus we get a nonzero value, 44.3, for the 3D distance, R_{15}. This differs from the value obtained for the stereoisomer, **6**, R_{16} = 46.9. The 3D abstract distances for the stereoisomer pairs (**5**,**6**) and (**7**,**8**) are 19.9 and 25.6, respectively. Thus, the 3D form of the molecular transform makes a clear distinction between stereoisomers that have, in general, quite different biological activities.

4. Isostructurality and Crystal Packing

Packing similarities among various crystal structures can be discussed in terms of *isostructurality* as proposed by Kálmán et al. [12]. If comparing unit cells *A* and *B* the isostructurality index is defined as follows

$$I_{AB}(n) = 100\times \left| [\Sigma(\Delta r^{i}_{AB})^2/n]^{1/2} - 1 \right| \quad (15)$$

where Δr^{i}_{AB} stands for the distance difference between crystal co-ordinates of the i-th pair of identical non-hydrogen atoms of the corresponding molecules of the asymmetric unit one in A the other in B. Summation includes n terms where n may vary between the number of atoms

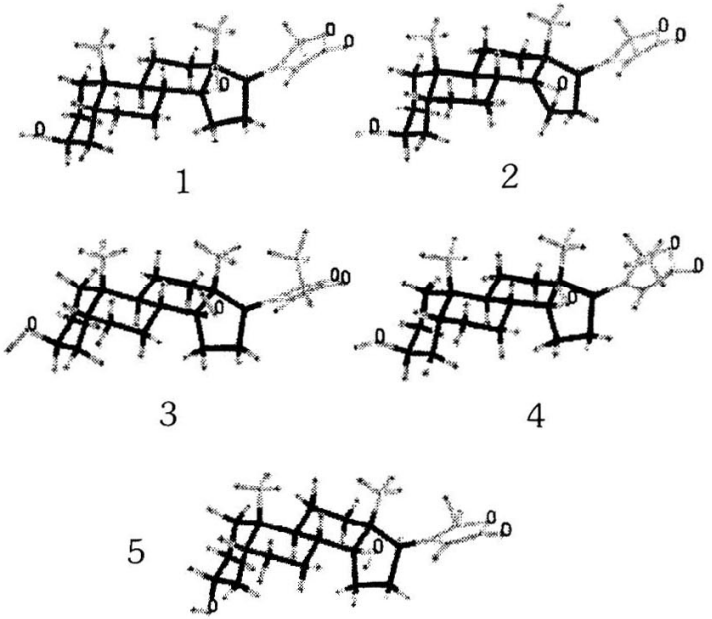

Figure 3. Molecular structures of the cardiotonic steroid derivatives for which we studied isostructurality. **1:** digitoxigenin, **2:** digirezigenin, **3:** 21*R*-methyldigitoxigenin, **4:** 21*S*-methyldigitoxigenin, **5:** 3-epidigitoxigenin. Figures drawn by the SYBYL software [15].

within an arbitrarily defined fragment and the maximum number of the common pairs of atoms. It is an important and interesting question, in which manner $I_{AB}(n)$ depends on R_{ab} calculated for the molecules a and b that build up the unit cells A and B.

Based on the work of Kálmán et al. [12] we investigated a series of cardiotonic steroids (cf. Figure 3). Considering the steroid ring structure with the fragment indicated by heavy lines (n = 17) and superimposing unit cell centres with cell edges parallel to each other, we calculated I_{AB} values for all possible pairs. The following linear regression equation was obtained (cf. Figure 4 for a plot)

$$I_{AB}(17) = -1.265R_{ab} + 88.9 \qquad r = 0.7935 \qquad n = 10 \tag{16}$$

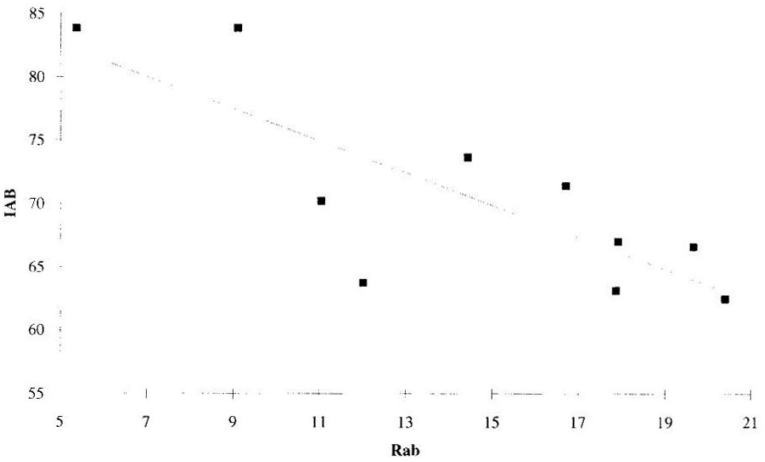

Figure 4. Graphical representation of eq. (16) for unit cells of crystals and free molecules of cardiotonic steroid derivatives depicted in Figure 3.

The correlation is fair, thus eq. (16) has an important meaning: *similar molecules tend to form similar unit cells in crystals*. This is represented graphically in Figure 5 where we compare pairs of similar and dissimilar molecules as well as the unit cells they form in the crystal.

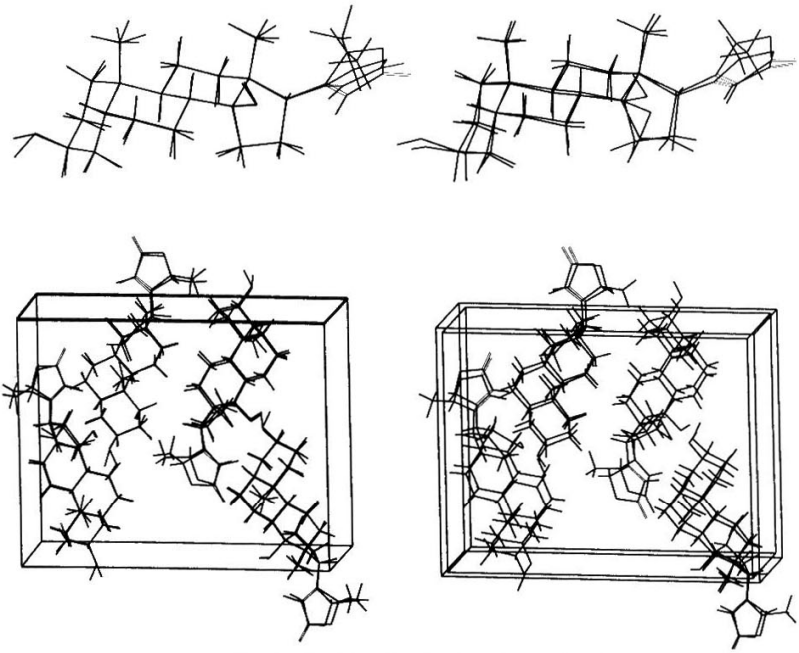

Figure 5. Superposition of molecules (above) and pairs of unit cells of crystals (below) of cardiotonic steroid derivatives. Left: **3-4** pair, R_{ab} = 5.37, $I_{AB}(17)$ = 83.9; right: **2-4** pair, R_{ab} = 20.40, $I_{AB}(17)$ = 62.5.

References

1. M.A. Johnson and G.M. Maggiora, eds., *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990.
2. P.G. Mezey, *Shape in Chemistry*, VCH Publishers, New York, 1993.
3. L.J. Soltzberg and C.L. Wilkins, *J. Am. Chem. Soc.* **98,** 7139 (1976).
4. R. Carbó, L. Leyda and M. Arnau, *Int. J. Quant. Chem.* **17,** 1185 (1980).
5. Z. Gabányi, P.R. Surján and G. Náray-Szabó, *Eur. J. Med. Chem.* **17,** 307 (1982).
6. P.G. Mezey, *Int. J. Quant. Chem. Quantum Biol. Symp.* **12,** 113 (1986).

7. E.E. Hodgkin and W.G. Richards, *Int. J. Quant. Chem. Quantum Biol. Symp.* **14,** 105 (1987).

8. G.A. Arteca and P.G. Mezey, *J. Phys. Chem.* **93,** 4746 (1989).

9. J.W. King, R.J. Kassel and B.B. King, *Int. J. Quant. Chem. Quantum Biol. Symp.* **17,** 27 (1990).

10. H. Lipson and C.A. Taylor, *Fourier Transforms and X-Ray Diffraction*, Bell, London, 1958.

11. I. Csorvássy, L. Tőzsér, L. Kárpáti and G. Náray-Szabó, *J. Math. Chem.***13,** 343 (1993).

12. A. Kálmán, L. Párkányi and G. Argay, *Acta Cryst.* **B49,** 1039 (1993).

13. L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research,* Academic Press, New York, 1976.

14. G. Náray-Szabó, *J. Mol. Struct. THEOCHEM* **134,** 401 (1986).

15. SYBYL Program Version 6.0a, TRIPOS Associates, St. Louis, MO, 1993.

16. R.S. Bohacek and C. McMartin, *J. Am. Chem. Soc.* **116,** 5560 (1994).