Journal of Computer-Aided Molecular Design, 3 (1989) 253–259
ESCOM

J-CAMD 067

A novel approach to molecular similarity

David L. Cooper[a,*] and Neil L. Allan[b]

[a] *Department of Chemistry, University of Liverpool, P.O. Box 147, Liverpool L69 3BX, U.K.*
[b] *Research and Technology Department, ICI Chemicals and Polymers Ltd., P.O. Box 8, The Heath, Runcorn, Cheshire WA7 4QD, U.K.*

Received 20 December 1988
Accepted 28 April 1989

Key words: Molecular similarity; Momentum space; Structure-activity relations; Drug design

SUMMARY

We review briefly the general problem of assessing the similarity between one molecule and another. We propose a novel approach to the quantitative estimation of the similarity of two electron distributions. The procedure is based on momentum space concepts, and avoids many of the difficulties associated with the usual position space definitions. Results are presented for the model systems $CH_3CH_2CH_3$, CH_3OCH_3, CH_3SCH_3, H_2O and H_2S.

INTRODUCTION

It would be very useful, particularly in the field of drug design, to be able to produce reliable answers to the general question: 'How similar are two molecules?'. This problem has received considerable attention from several research groups. However, the concept of molecular similarity is sufficiently loose and imprecise that a number of very different approaches has been proposed.

Some research groups search databases for given pharmacophoric patterns [1] in order to establish series of 'similar' molecules. Others use molecular graphics to overlay active molecules in an attempt to characterize the binding site into which 'similar' molecules must also fit. However, it is not at all clear how to superimpose molecules which appear to be very different: these approaches over-emphasize the bonding topology. To deal with this particular problem, methods have been developed to obtain matches between 'molecular accessible surfaces' or between molecular surfaces onto which properties such as the electrostatic potential have been mapped [2].

Ultimately, the chemistry of a molecule must depend to a very large extent on its electronic

*To whom correspondence should be addressed.

structure. Not surprisingly, an approach that is attracting considerable current interest in the pharmaceutical and agrochemical industries focuses on the similarity between different electron distributions. The usual approach is to use quantum chemical calculations to examine the overlap between the electron densities of different molecules. In this paper we propose a novel, but very simple, alternative strategy which circumvents many of the technical and conceptual problems associated with the evaluation of such overlaps. Instead of the usual position space representation, we examine the similarity between electron densities which are functions of the *momenta* rather than of the positions of the electrons. In this way, we concentrate more on the outer valence-electron density.

Carbó et al. [3] have introduced a similarity index r_{AB} which defines the similarity between two molecules in terms of the position space overlap of their charge densities $\rho(\boldsymbol{r})$,

$$r_{ab} = \frac{\int \rho_A(\boldsymbol{r})\rho_B(\boldsymbol{r})d\boldsymbol{r}}{(\int \rho_A^2(\boldsymbol{r})d\boldsymbol{r})^{1/2}(\int \rho_B^2(\boldsymbol{r})d\boldsymbol{r})^{1/2}} \tag{1}$$

and much use of this has been made by Richards and co-workers [4–6], and by Ponec [7, 8]. This definition of similarity is much more sensitive to the shapes of the electron densities than to their overall magnitudes. A detailed account of molecular orbital similarity measures has been presented by Carbó and Domingo [9], who have also given computational formulae for the four-centre, overlap-like integrals which arise for molecular orbital theory wave functions constructed from Gaussian functions.

A slight variation on this definition of molecular similarity has been proposed [10]:

$$s_{AB} = \frac{2\int \rho_A(\boldsymbol{r})\rho_B(\boldsymbol{r})d\boldsymbol{r}}{(\int \rho_A^2(\boldsymbol{r})d\boldsymbol{r}) + (\int \rho_B^2(\boldsymbol{r})d\boldsymbol{r})} \tag{2}$$

This formulation of similarity is much more sensitive to the magnitudes of $\rho(\boldsymbol{r})$ than is r_{AB}. Both r_{AB} and s_{AB} have also been used with electrostatic potentials [10, 11], rather than with electron densities. An interesting summary of such approaches, and of possible future applications, has appeared recently [12].

Evaluation of the indices r_{AB} and s_{AB} requires an arbitrary decision as to how to superimpose the two molecules A and B, and the values are very sensitive to the relative positions (and orientations). One solution is to adjust the positions of the two molecules until the indices are maximized, but this can be expensive [10]. The indices are especially sensitive to the positions of heavy atoms because of the concentration of electron density near the nucleus. Furthermore, if we compare two molecules which differ only by a small displacement of one heavy atom, the position of this one nucleus is likely to dominate the similarity index. Although 'valence only' comparisons do alleviate some of these difficulties, and there have been some successful applications, the basic underlying problem remains: the indices r_{AB} and s_{AB} tend to be more sensitive to the positions of the nuclei than to the long-range valence-electron density. Given two molecules with identical long-range valence-electron densities, but different nuclear coordinates and different core-electron distributions, a measure of similarity which returns a value close to 100% is likely to be more successful in attempts to correlate structure with activity. With this in mind, we investigate indices based on momentum space quantities.

MOMENTUM SPACE ELECTRON DENSITIES

The momentum space wave function is the Fourier transform of that in position space. The transformation preserves direction so that, for example, it is possible to distinguish components of the total momentum parallel or perpendicular to particular bonds or to planes of symmetry.

Consider a molecular orbital $\psi(\boldsymbol{r})$ formed by the overlap of atomic basis functions $\varphi_\alpha(\boldsymbol{r})$ centred on nuclei at positions $\boldsymbol{R}_\alpha$:

$$\psi(\boldsymbol{r}) = \sum_\alpha c_\alpha \varphi_\alpha(\boldsymbol{r}) \tag{3}$$

The corresponding momentum space orbital $\psi(\boldsymbol{p})$ takes the form

$$\psi(\boldsymbol{p}) = \sum_\alpha c_\alpha \varphi_\alpha(\boldsymbol{p}) \exp(-i\boldsymbol{p}.\boldsymbol{R}_\alpha) \tag{4}$$

where the $\varphi_\alpha(\boldsymbol{p})$ are the Fourier transforms of the $\varphi_\alpha(\boldsymbol{r})$, i.e.

$$\varphi_\alpha(\boldsymbol{p}) = (2\pi)^{-3/2} \int \varphi_\alpha(\boldsymbol{r}) \exp(-i\boldsymbol{p}.\boldsymbol{r}) d\boldsymbol{r} \tag{5}$$

The contribution of molecular orbital $\psi(\boldsymbol{p})$ to the total $\boldsymbol{p}$-space density $\rho(\boldsymbol{p})$ is simply the square modulus, $|\psi(\boldsymbol{p})|^2$.

The information about the nuclear positions $\boldsymbol{R}_\alpha$ is contained solely in the exponential 'phase factors' (Eq. 4) which may produce striking oscillatory structure in individual delocalized orbitals, but which tend to have relatively little effect on the total density $\rho(\boldsymbol{p})$. We have demonstrated in previous work on the momentum densities of small molecules [13,14] that the $\boldsymbol{p}$-space representation can be used to highlight the process of bond formation in a particularly striking way. For large molecules, such as long polyenes [15], we have shown that the momentum space formalism is particularly suitable for comparing the electron distributions of systems with different nuclear frameworks.

We now define a similarity index in terms of the momentum densities $\rho(\boldsymbol{p})$ of the two molecules. For reasons we discuss later, it might be useful to introduce powers of $p = |\boldsymbol{p}|$ into the integrands. By analogy with Eq. 2, we define $S_{AB}(n)$ according to

$$S_{AB}(n) = \frac{2\int p^n \rho_A(\boldsymbol{p}) \rho_B(\boldsymbol{p}) d\boldsymbol{p}}{(\int p^n \rho_A^2(\boldsymbol{p}) d\boldsymbol{p}) + (\int p^n \rho_B^2(\boldsymbol{p}) d\boldsymbol{p})} \tag{6}$$

The general shapes of the momentum densities for different molecules are fairly similar [16], and so we would not expect the momentum space analogue of r_{AB} (Eq. 1) to be a very discriminating measure of similarity.

The index $S_{AB}(n)$ depends only on the relative orientation of the two molecules and not on the distance between them in $\boldsymbol{r}$-space: most of the problems connected with superimposing the two molecules are circumvented. Small displacements of heavy atoms in $\boldsymbol{r}$-space are likely to cause only very small changes in $\boldsymbol{p}$-space.

Low values of $\boldsymbol{p}$ correspond to the slowly varying outer valence-electron density in $\boldsymbol{r}$-space. The

momentum density decays rapidly with p and thus is dominated by low values of the momentum. Whereas the form of the electron density in r-space is dominated to a large extent by the core electrons, and thus by the positions of the nuclei, $\rho(\boldsymbol{p})$ highlights the chemically more interesting features of the valence-electron distribution.

RESULTS FOR MODEL SYSTEMS

In order to demonstrate the method we have considered the model systems $CH_3CH_2CH_3$, CH_3OCH_3, CH_3SCH_3, H_2O and H_2S. The first three of these molecules have been considered in various investigations of molecular similarity [4, 5, 10]. Note that the replacement of $-CH_2-$ by -S- is a very common device in drug design in that it often leads to very similar activity: the same is not generally true for -O-. An advantage of this series of molecules is that it is particularly straightforward to recognize appropriate relative orientations. For each molecule, we take the C_2-axis to be the z-direction, with the y-axis in the σ_v' plane.

Self-consistent field (SCF) wave functions were obtained for the equilibrium geometries of these molecules using the GAMESS [17] program with a 4-31G spherical Gaussian basis set [18]. The Fourier transforms of Gaussian functions have simple analytic forms [19], and so the evaluation of $\rho(\boldsymbol{p})$ is no more difficult than that of $\rho(\boldsymbol{r})$. Although it would be straightforward here to use larger basis sets and to include some of the effects of electron correlation, the calculation of more sophisticated wave functions would probably be too expensive for the larger systems of real interest.

In Fig. 1 we show the total momentum space density $\rho(\boldsymbol{p})$ calculated for H_2O in the plane $p_x = 0$. In position space, the molecule is oriented so that all the nuclei lie in the plane $x = 0$. In addition to an inversion centre, which is linked to the absence of net translational motion, $\rho(\boldsymbol{p})$ must exhibit the same symmetry elements as $\rho(\boldsymbol{r})$.

We have calculated values of $S_{AB}(0)$ for each pair of molecules. For our model systems, the values of $S_{AB}(0)$ range from 47.7% to 99.8%. We have also evaluated the momentum space analogue

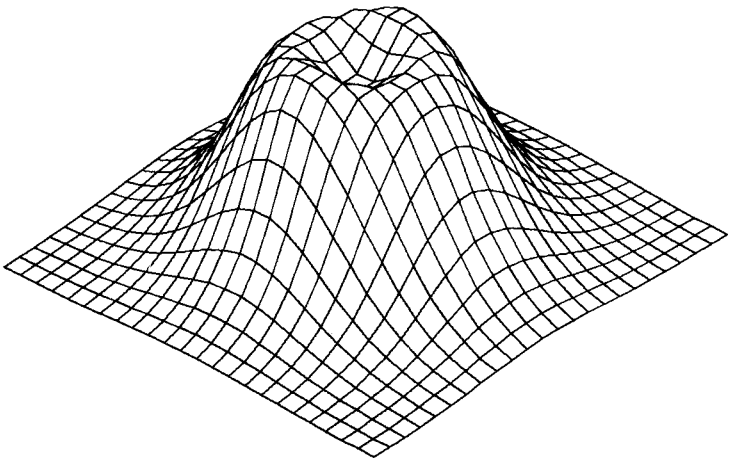

Fig.1. A representation of the total momentum density $\rho(\boldsymbol{p})$ calculated for H_2O in the plane $p_x = 0$. The nuclei lie in the plane $x = 0$ in position space.

of r_{AB}: this index was found to show exactly the same trends but was very much less discriminating, as expected, with a range of values from 93.8% to 99.9%.

It is, of course, unreasonable to expect a single number such as $S_{AB}(0)$ (or s_{AB}) to indicate whether or not two electron densities $\rho(\boldsymbol{p})$ (or $\rho(\boldsymbol{r})$) are genuinely very similar over the entire range of $\boldsymbol{p}$ (or $\boldsymbol{r}$). By introducing different powers of p into the integrands, the indices $S_{AB}(n)$ emphasize different regions of the electron distribution. When considering the question of similarity between different molecules, it is important to consider a range of values of n. For our series of model systems, values of $S_{AB}(n)$ for $n = -1, 0, 1, 2$ are reported in Table 1.

We would expect high similarity indices between all three molecules of general formulae $(CH_3)_2X$, and this is indeed found to be the case. It is particularly gratifying that the highest values of $S_{AB}(n)$, for all n, occur in the comparison of $(CH_3)_2S$ and $(CH_3)_2CH_2$. For the comparisons of these molecules, the values of $S_{AB}(n)$ are virtually independent of n.

It is not unreasonable that the least similar pair of molecules, according to $S_{AB}(n)$, consists of H_2O and CH_3SCH_3, with values between 43.9% and 54.5%. The variation with n shows that the momentum densities are least similar at low values of p. Presumably because of the non-bonding electrons on sulphur, H_2S and CH_3SCH_3 are most similar at low p.

The only puzzling feature of Table 1 is the high set of values of $S_{AB}(n)$ for H_2S and CH_3OCH_3.

TABLE 1
MOMENTUM SPACE SIMILARITY INDICES $S_{AB}(n)$*, EXPRESSED AS PERCENTAGES, FOR $n = -1, 0, 1, 2$

	n	CH_3OCH_3	CH_3SCH_3	$CH_3CH_2CH_3$	H_2O	H_2S
CH_3OCH_3	−1	100	97.4	98.4	53.3	86.1
	0	100	97.5	98.5	57.3	83.3
	1	100	97.5	98.6	60.5	79.7
	2	100	97.3	98.8	63.3	76.8
CH_3SCH_3	−1		100	99.8	43.9	76.8
	0		100	99.8	47.7	75.1
	1		100	99.6	51.3	73.4
	2		100	98.8	54.5	72.9
$CH_3CH_2CH_3$	−1			100	45.7	78.9
	0			100	49.9	77.3
	1			100	53.7	75.1
	2			100	57.7	73.7
H_2O	−1				100	76.4
	0				100	81.0
	1				100	84.6
	2				100	85.7
H_2S	−1					100
	0					100
	1					100
	2					100

*$S_{AB}(n)$ is defined in Eq. 6.

At least in momentum space, for the values of n considered here, H_2S appears to be more similar to CH_3OCH_3 than to CH_3SCH_3. However, all of the values of $S_{AB}(n)$ show significant variation with n when one of the partners is H_2O or H_2S. For $n \geq 1$, the values of $S_{AB}(n)$ indicate that H_2S is more similar to H_2O than to any of the $(CH_3)_2X$ molecules. As suggested earlier, molecules should only be considered to be 'similar' if several indices have high values, with each sampling different regions of the electron distributions.

For the three $(CH_3)_2X$ systems, values of r_{AB} and s_{AB} have been published based on ***r***-space electron densities [5], electrostatic potentials [10], and molecular electric fields [10]. Only calculations of r_{AB} using valence-electron densities [5] reproduced the order observed in Table 1. The s_{AB} results based on electrostatic potentials [10] were particularly surprising in that they suggested a very low similarity between propane and thioether (3%), but a very high similarity between ether and thioether (92%). Nevertheless, electrostatic potentials have been widely used for studies of structure–activity relationships, albeit mostly in terms of comparison by eye of colour-coded surfaces rather than by quantitative measures.

CONCLUSIONS

We have proposed a new formulation of molecular similarity, based on momentum space. This approach avoids many of the difficulties associated with position space definitions, and the method is not computationally expensive. In the present work, we have considered only a series of very simple model compounds, but it would not be difficult to apply the method to larger molecules. In cases where the preferred relative orientation of the molecules is not self-evident, if would be very straightforward to maximize any of the $S_{AB}(n)$.

For comparisons of larger molecules which differ only in one or two sites, it would probably be more useful to transform to ***p***-space only that part of the ***r***-space wave function which describes the functional groups of interest. A very interesting direction for future work would be to use momentum space concepts such as $S_{AB}(n)$ to compare HOMOs and LUMOs of different molecules, and to correlate these with reactivity.

It is important to recognize that the propensity of a particular molecule to undergo different chemical reactions depends on many different features. No single index can be expected to be a universal panacea. In a 'real' application it would be more useful to examine simultaneously a variety of measures of similarity, each emphasizing different aspects of the molecular shape, bonding topology and electronic structure effects.

It is very much hoped that the quantitative measures described in this work will play a useful role in attempts to relate molecular structure to biological activity and chemical reactivity.

REFERENCES

1 Brint, A.T. and Willett, P., J. Comput.-Aided Mol. Design, 2 (1988) 311–320, and references therein.
2 Dean, P.M., Callow, P. and Chau, P.-L., J. Mol. Graph., 6 (1988) 28–34, and references therein.
3 Carbó, R., Leyda, L. and Arnau, M., Int. J. Quantum Chem., 17 (1980) 1185–1189.
4 Bowen-Jenkins, P.E., Cooper, D.L. and Richards, W.G., J. Phys. Chem., 89 (1985) 2195–2197.
5 Bowen-Jenkins, P.E. and Richards, W.G., Int. J. Quantum Chem., 30 (1986) 763–768.
6 Richards, W.G., Pure and Appl. Chem., 60 (1988) 277–279.
7 Ponec, R., Coll. Czech. Chem. Commun., 52 (1987) 555–562.

8 Ponec, R., Z. Phys. Chem. Leipzig, 268 (1987) 1180–1188.
9 Carbó, R. and Domingo, Ll., Int. J. Quantum Chem., 32 (1987) 517–545.
10 Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem.: Quantum Biology Symp., 14 (1987) 105–110.
11 Nakayama, A. and Richards, W.G., Quant. Struct. Act. Relatsh., 16 (1987) 153–157.
12 Richards, W.G. and Hodgkin, E.E., Chem. Br., 24 (1988) 1141–1144.
13 Cooper, D.L. and Allan, N.L., J. Chem. Soc., Faraday Trans. 2, 83 (1987) 449–460.
14 Allan, N.L. and Cooper, D.L., J. Chem. Soc., Faraday Trans. 2, 83 (1987) 1675–1687.
15 Cooper, D.L., Allan, N.L. and Grout, P.J., J. Chem. Soc., Faraday Trans. 2, 85 (1989) 1519–1529.
16 See, for example, Rawlings, D.C. and Davidson, E.R., J. Phys. Chem., 89 (1985) 969–974.
17 Guest, M.F. and Kendrick, J., GAMESS User Manual, CCP1/86/1, Daresbury Laboratory, SERC, 1986.
18 Frisch, M.J., Pople, J.A. and Binkley, J.S., J. Chem. Phys., 80 (1984) 3265–3269, and references cited therein.
19 Kaijser, P. and Smith, V.H., Adv. Quantum Chem., 10 (1977) 37–76.