A Semi-empirical Method for Calculating Molecular Similarity

Edward E. Hodgkin and W. Graham Richards

Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, U.K.

An approximate method for calculating molecular similarity has been developed, based on the notion of standard electron densities for molecular fragments; the technique should provide a means of comparing large biological molecules.

The substitution of one group for another in a molecule is used widely in the pharmaceutical industry to try to improve or reproduce biological activity. The concept of bioisosterism[1] is useful in predicting how effective a particular substitution might be.

The similarity of two molecules, A and B, may be quantified theoretically by comparing their electron densities, ρ_A and ρ_B and calculating an index of similarity, R_{AB}, equation (1), as first introduced by Carbo.[2]

$$R_{AB} = \frac{\int \rho_A \rho_B dv}{(\int \rho_A{}^2 dv)^{1/2} (\int \rho_B{}^2 dv)^{1/2}} \qquad (1)$$

The parameter R_{AB} takes values in the range 0 to 1. A value of one is obtained when comparing identical molecules.

An *ab initio* method for the computation of the similarity index has been developed.[3] However, this approach requires the calculation of a large number of four-centre integrals and is costly in computer time. Hence there is a need for a fast approximate method for calculating the molecular similarity.

The method presented here is based on the idea that a molecule can be divided into a number of typical fragments. For example, we assume that every bond between two sp^3 carbon atoms looks like the one in ethane. Thus, characterization of the electron density of a series of simple molecules enables one to build a description of the electron density in large complex molecules.

The electron density used to calculate the similarity index is represented by a series of spherical Gaussian functions, of the form $G = ce^{-\alpha r^2}$. The Gaussians are parameterized by matching with an *ab initio* density distribution obtained using the *ab initio* molecular orbital program Gaussian 80[4] with an STO-3G basis set and DENPOT,[5] a program to calculate electron density values from an *ab initio* wavefunction. The standard values of the parameters c and α are obtained by performing a least-squares fit.[6] The positions of the Gaussian functions are determined by the positions of the maxima in the *ab initio* electron density.

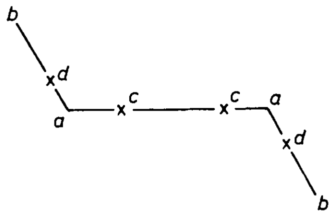

Figure 1. Positions of Gaussian functions in ethane. Distance *ac* is 0.38 Å, *ad* is 0.37 Å. The molecule is viewed in the HCCH plane of the staggered conformation.

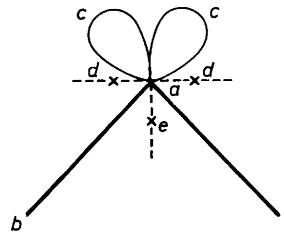

Figure 2. Positions of Gaussian functions in dimethyl ether. The molecule is viewed in the COC plane, with the plane *cac* perpendicular to the page. Distance *ac* is 0.25 Å, *ad* is 0.24 Å and *ae* is 0.24 Å. Angle *cac* is 152°.

Table 1. Parameterization of Gaussian functions for ethane.

Position[a]	Gaussian exponent	Proportionality constant
a	179.0[b]	68.32
	1.39	0.24
	26.66	−0.27
b	6.46	0.30
c	8.28	0.03
d	23.37	0.03

[a] The positions correspond to those in Figure 1. [b] 1s core.

We have obtained standard Gaussians to describe alkanes and ethers, the molecules studied being ethane and dimethyl ether. The core and valence electron density are treated separately as the former masks the structure of the latter. The positions of the Gaussians used for ethane and dimethyl ether are shown in Figures 1 and 2. The corresponding Gaussian exponents and proportionality constants are shown in Tables 1 and 2.

In the case of ethane, the Gaussians were centred on the maxima in the *ab initio* electron density distribution. Maps of the difference between *ab initio* and Gaussian approximation electron densities of ethane showed that this approach was satisfactory for ethane. Dimethyl ether proved to be more difficult. The sp^3 lobes and lone-pairs of the oxygen atom do not appear as separate maxima in the electron density distribution. However, on fitting Gaussians to the maxima in the *ab initio* density distribution, density difference maps showed the positions where further Gaussians should be added to improve the approximation of the *ab initio* electron density.

A series of bioisosteres, Me_2CH_2, Me_2O, Me_2S, has previously been studied at the *ab initio* level.[7] Thus, we have been able to evaluate the Gaussian approximation method by comparing propane with dimethyl ether. The similarity index was calculated as a function of the separation between the centroids of the molecules for both the total electron density and the valence electron density.

Table 2. Parameterization of Gaussian functions for dimethyl ether.

Position[a]	Gaussian exponent	Proportionality constant
a	295.0[b]	162.0
	4.06	1.54
	77.58	−1.36
b	179.0[b]	68.32
	1.39	0.24
	26.66	−0.27
c	17.03	0.26
d	16.36	−0.34
e	11.76	−0.38

[a] The positions correspond to those in Figure 2. [b] 1s core.

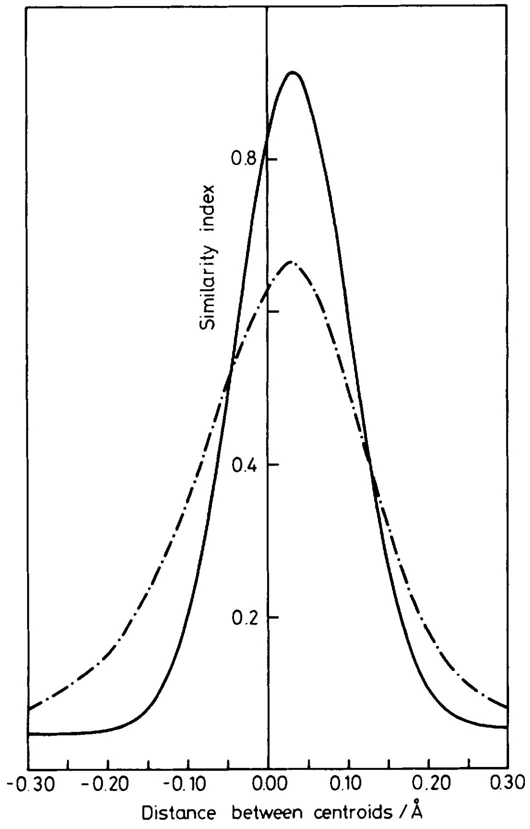

Figure 3. Similarity index of total electron density for the comparison of Me_2CH_2 with Me_2O: ——— Gaussian approximation; –·–·– *ab initio*.

The results are shown in Table 3 and in Figures 3 and 4. It can be seen that agreement between *ab initio* and Gaussian approximation values is extremely good for the valence electron density comparisons, but less good for the total electron density. Note that the scale in Figure 4 is ten times more discriminating than that in Figure 3. However, the

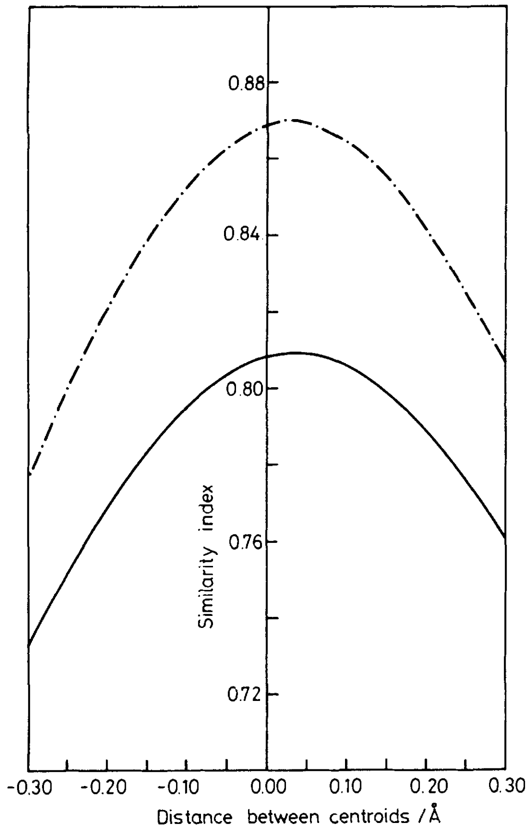

Figure 4. Similarity index of valence electron density for the comparison of Me_2CH_2 with Me_2O: ——— Gaussian approximation; –·–·– *ab initio*.

valence electron density is of greater importance in prediction of chemical activity.[7] In addition, when using the similarity index as a criterion for the best superposition of the molecules both the total electron density and the valence electron density give relative positions of the two molecules which are in agreement with the results of full *ab initio* computations.

Application of the new method to other systems requires the parameterization to be extended to more bond types. It is hoped that the technique will be used to study large molecules of biological interest, such as the ring system in prostaglandins. A pattern of similarity values might be observed for such systems, which may be related to biological activity.

Table 3. Values of the similarity index for the comparison of propane with dimethyl ether.

	Full		Valence		
Method of matching[a]	*Ab initio* R_{AB}	Gaussian R_{AB}	*Ab initio* R_{AB}	Gaussian R_{AB}	Δ[b]
(1)	0.61	0.89	0.87	0.81	0.05
(2)	0.63	0.83	0.87	0.81	0.00

[a] The matching methods are: (1) the central atoms are superimposed; (2) the molecule centroids are superimposed. In each case the molecules have a common symmetry axis. [b] Δ Represents the distance between the centroids of the two molecules in (1).

It should be noted that the approximate method described here represents a considerable saving in time compared with the *ab initio* method. A single computation of the similarity index for propane and dimethyl ether requires about two seconds of central processor time on a VAX 11/780 computer.

The authors thank Dr. M. M. Hann of Glaxo Group Research Ltd., for helpful discussions.

Received, 11th June 1986; Com. 799

References

1 C. W. Thornber, *Chem. Soc. Rev.*, 1979, **8**, 563.
2 R. Carbo, L. Leyda, and M. Arnau, *Int. J. Quant. Chem.*, 1980, **17**, 1185.
3 P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.*, 1985, **89**, 2195.
4 J. S. Binkley, R. A. Whiteside, R. Krishnan, R. Seeger, D. J. de Frees, H. P. Schlegel, S. Topiol, L. R. Khan, and J. A. Pople, *Quantum Chemistry Program Exchange*, 1981, **13**, 406.
5 D. Peeters and M. Sana, *Quantum Chemistry Program Exchange*, Program QCPE 360.
6 P. E. Gill and W. Murray, *SIAM J. Numerical Anal.*, 1978, **15**, 1977.
7 P. E. Bowen-Jenkins and W. G. Richards, *J. Chem. Soc., Chem. Commun.*, 1986, 133.