Journal of Computer-Aided Molecular Design, 5 (1991) 427–439
ESCOM



J-CAMD 140

# Similarity of molecular shape

Amatzya Y. Meyer* and W. Graham Richards

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

Received 31 August 1990
Accepted 24 July 1991

*Key words:* Similarity; Molecular shape; Chirality

## SUMMARY

The similarity of one molecule to another has usually been defined in terms of electron densities or electrostatic potentials or fields. Here it is expressed as a function of the molecular shape. Formulations of similarity (S) reduce to very simple forms, thus rendering the computerised calculation straightforward and fast. 'Elements of similarity' are identified, in the same spirit as 'elements of chirality', except that the former are understood to be variable rather than present-or-absent. Methods are presented which bypass the time-consuming mathematical optimisation of the relative orientation of the molecules. Numerical results are presented and examined, with emphasis on the similarity of isomers. At the extreme, enantiomeric pairs are considered, where it is the dissimilarity (D = 1 − S) that is of consequence. We argue that chiral molecules can be graded by dissimilarity, and show that D is the shape-analog of the 'chirality coefficient', with the simple form of the former opening up numerical access to the latter.

## INTRODUCTION

The major discriminating force between different molecules that bind to a single receptor is repulsion. This is another way of saying that the most important feature is fitting in the simple geometrical sense: the power of the distance in any force field is much higher on the repulsive side of the potential than any attraction. Quantitative measures of molecular similarity have been used extensively in structure–activity correlations, in particular those measures that are based on comparisons of electrostatic potential and field. These relate to attractive forces between the small molecule and its receptor; respectively due to ionic or dipolar forces. Here we present a method of comparing molecules based strictly on shape. The method permits optimisation of similarity and,

* The late Amatz Meyer was formerly of the Department of Organic Chemistry, Hebrew University, Jerusalem 91904, Israel.

0920-654X/$ 5.00 

most important, calculations are very fast. We believe that the method presented here would permit a company to compare a molecule with every member of a database of hundreds of thousands of compounds in a reasonable time on a small computer.

## METHOD

Two lines of study converge towards the present work. One is the analysis of molecular similarity and dissimilarity. In 1980, Carbo et al. proposed a measure of similarity: the two molecules to be compared are computationally superposed, and their similarity S is expressed in terms of the charge-density distributions [1] ($\rho_1$ and $\rho_2$, Eq. 1). Then, Richards and Hodgkin noticed that S is not sensitive to dissimilarity of intensities: distributions ($\rho_1$, $\rho_2$) and ($\rho_1$, $n\rho_2$) lead to the same S [2]. They proposed an alternative index, which is free from this drawback ($S_H$, Eq. 2). Very recently, Gilat [3] and ourselves [4] studied the gradation of molecular chirality. It is shown below that Gilat's 'chiral coefficient' $\chi$ coincides with $1 - S$, a measure of molecular dissimilarity.

$$S = \frac{\int \rho_1 \rho_2 \, dV}{(\int \rho_1^2 dV)^{1/2} (\int \rho_2^2 dV)^{1/2}} \tag{1}$$

$$S_H = \frac{2 \int \rho_1 \rho_2 dV}{\int \rho_1^2 dV + \int \rho_2^2 dV} \tag{2}$$

The second line of study is the assessment of molecular size and shape [5] via point-counting algorithms [6,7]. Each atom in a molecule is circumscribed by a sphere of a typical radius, whereupon a representation is obtained of the molecule as a system of interlocking spheres. This body is computationally enclosed in a gridded box, and the mesh points are scanned. Counts are obtained of points that occur within the body, points close to its surface, to given planes, etc. From the counts, attributes such as the molecular volume, surface area and cross-sectional areas, are evaluated and used to obtain compounded attributes. To emphasise the simplicity of the procedure and for later reference, we cite two ways to estimate the molecular volume V (Eq. 3). Here g is the grid-increment (in units of length), T and $T_B$ are, respectively, the total number of points within the molecule and within the box, and $V_B$ is the box volume.

$$V = g^3 T; \quad V = TV_B/T_B \tag{3}$$

Now we go a step further and report on assessing similarities by the point-counting pathway, and on the results obtained in some examples. With respect to density-distribution procedures, one may note two advantages and one drawback. One advantage is that formulae involve very simple quantities, so that relations between S, $S_H$ and $\chi$ stand out clearly. Another is that computation demands very little in human effort and computer time. The disadvantage, of course, is that a grid-point either contributes or does not contribute to S, whereas $\rho_1 \rho_2$ at the point may take any value within a range: unlike the charge-density pathway, point-counting cannot respond to the gradation of electronic density.

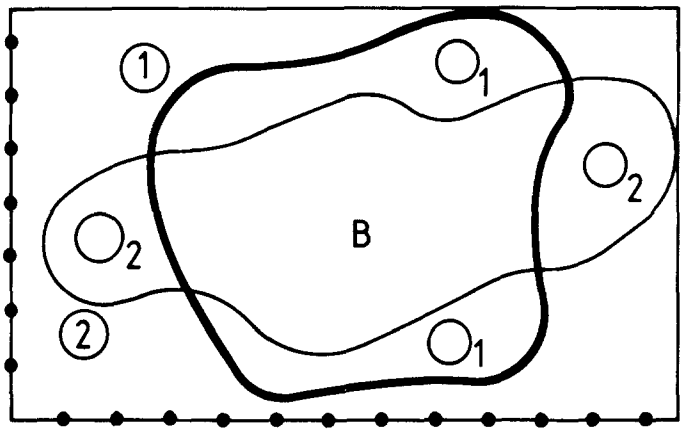


Fig. 1. Schematic representation of a cut through two superposed molecules and the containing box.

*Point-counting formulae*

Consider two molecules, 1 and 2, each defined as a body in space by the coordinates and radii of the constituent atoms. The molecules are computationally slid onto each other until proper overlap is achieved, and the resulting body is enclosed in a box of grid points (Fig. 1). In practice, we use molecular-mechanical geometries for aliphatics (MM2 [8] for hydrocarbons, other force fields [9] for substituted hydrocarbons) and literature geometries (measured or other) for aromatics; van der Waals atomic radii [6]; and a grid of 0.02 nm. Finer grids prolong the time of computation without affecting the results significantly. Points in the box are scanned, and count is obtained of:

- $O_1$ and $O_2$, number of points occurring, respectively, only in molecule 1 or only in molecule 2;
- B, number of points in both molecules;
- Total number of points in each molecule, $T_1 = O_1 + B$ and $T_2 = O_2 + B$.

With these numbers, the point-counting analogs of Eqs. 1 and 2 take the form of Eqs. 4 and 5. S and $S_H$ take values in the range 0 to 1.

$$S = \frac{B}{(T_1 T_2)^{1/2}} \tag{4}$$

$$S_H = \frac{2B}{T_1 + T_2} \tag{5}$$

Compilations of charge-density similarities [10] (Eqs. 1 and 2) reveal that, for a given pair of molecules, $S > S_H$. Eqs. 4 and 5 are simple enough to allow a proof. It is based on the quotient

$$Q = \frac{S}{S_H} = \tfrac{1}{2}\left(\frac{T_1}{T_2}\right)^{\frac{1}{2}} + \left(\frac{T_2}{T_1}\right)^{\frac{1}{2}}$$

Case 1 (the more common): $T_1 \neq T_2$. Q is half the sum of a number and its reciprocal. Such a sum is always larger than 2, so that $Q > 1$ and $S > S_H$.

Case 2: $T_1 = T_2$ (identical molecules or enantiomers, cf. Eq. 3) or $T_1 \sim T_2$ (isomers of position, diastereoisomers). $Q = 1$ or $Q \sim 1$, hence $S = S_H$ or approximately so, and one may put $T = T_1 \sim T_2$. Since $T_i = O_i + B$ ($i = 1,2$), we conclude $O = O_1 \sim O_2$, and the index of similarity becomes as in Eq. 6.

$$S \text{ (for isomers)} = \frac{2B}{T_1 + T_2} = \frac{B}{O + B} \tag{6}$$

$$D = 1 - S = \frac{O}{O + B} \text{ (for isomers)} \tag{7}$$

*Dissimilarity*

For any pair of molecules, it is convenient to define the dissimilarity as $D = 1 - S$. If the molecules are isomeric (excepting, perhaps, functional isomers), D takes the form in Eq. 7. Note that enantiomers and other isomers are characterised by $O_1 \sim O_2$, not by $S \sim 1$: enantiomers can be quite dissimilar but must have $O_1 \sim O_2$, while non-isomeric molecules can be quite similar but must have $O_1 \neq O_2$. For example, the similarity of the two enantiomers of CHFClBr reaches only to $S = 0.90$, while the similarity of nonane to decane amounts to 0.95. In the former case the counts at $g = 0.02$ nm are $O_1 = 854$, $O_2 = 857$ and $B = 7636$; in the latter, these are $O_1 = 45$, $O_2 = 2145$ and $B = 20\,386$. It is for a technical reason that $O_1$ comes out a bit different from $O_2$ in the case of enantiomers. As the algorithm runs, the containing box (Fig. 1) is traced such that its origin coincides with the lower extremities of the body contained. Hence, the two enantiomers are confined to one octant, and not placed symmetrically with respect to a coordinate plane.

*Elements of similarity*

For a given pair of molecules there is an infinite set of mutual orientations, each associated with its own indices of similarity. Investigators have aimed at positioning the molecules so as to obtain the highest possible value of S. Whatever the philosophy behind this choice, it certainly standardises computations by unambiguously singling out one particular solution from many. In practice, one uses an optimising routine that shifts one molecule about the other in a way to maximise S [11]. Since the procedure is not economical in computer-time, we sought short cuts to satisfactory superpositions. Before describing our procedure, it is convenient to consider 'elements of similarity', viz. traits that two molecules may have in common. Four types come to mind.

(a) Atomic, i.e., the identity of atoms and chemical groups. A methyl group in one molecule is similar and approximately superposable on a methyl group in a different molecule.

(b) Molecular, i.e., similarity of bond lengths, valence angles or torsional angles. For example, bond Br-Cl in the interhalogen BrCl (0.214 nm) is close in length to bond C-I (0.215 nm) in alkyl iodides. The two bonds are superposable, thus bringing the atomic nuclei at their terminals into overlap.

(c) Symmetry, i.e., symmetry elements that characterise both molecules. Dimethyl ether and *meta*-xylene have both an axis and two planes of symmetry, which constitute three superposable elements of similarity.

(d) Cluster. In some molecules the atoms cluster about an axis or about a plane. For example, in the chair-form of cyclohexane atoms cluster about the 'average skeletal plane of the ring'.

This plane is distinct from, and perpendicular to the planes of symmetry. If both the molecules compared have a cluster-axis or a cluster-plane, then this axis or plane constitutes an element of similarity.

When orienting molecules with a view to calculating S, one would wish to superpose as many elements as possible. Clearly, not all similarity elements can be brought simultaneously into coincidence. Compromises are unavoidable, and one of the compromises should be better than others, that is, lead to a higher B, lower O's and a higher S (cf. Eqs. 4 and 5).

*Example*

Suppose one has to compare axial with equatorial methylcyclohexane. Out of the infinity of conceivable orientations, two stand out: in one, the rings are superposed or approximately so (Fig. 2A); in the other, the H-C-Me moieties are superposed (Fig. 2B). In the former, numerous elements coincide or almost so (16 atoms, all C-C bonds but one, all C-H bonds but four, most valence and dihedral angles, the planes of symmetry and of cluster); in the latter, only a few. Clearly, rigorous mathematical optimisation is not required to single out the situation in Fig. 2A as the best of all.

*Procedure*

Two molecules, each positioned arbitrarily, have to be brought into efficient superposition. Rotation-translation of each into its principal (inertial) coordinate system [12] is the procedure that comes first to mind. This is a 'transformation-by-mass': each atom is attributed a weight equal to

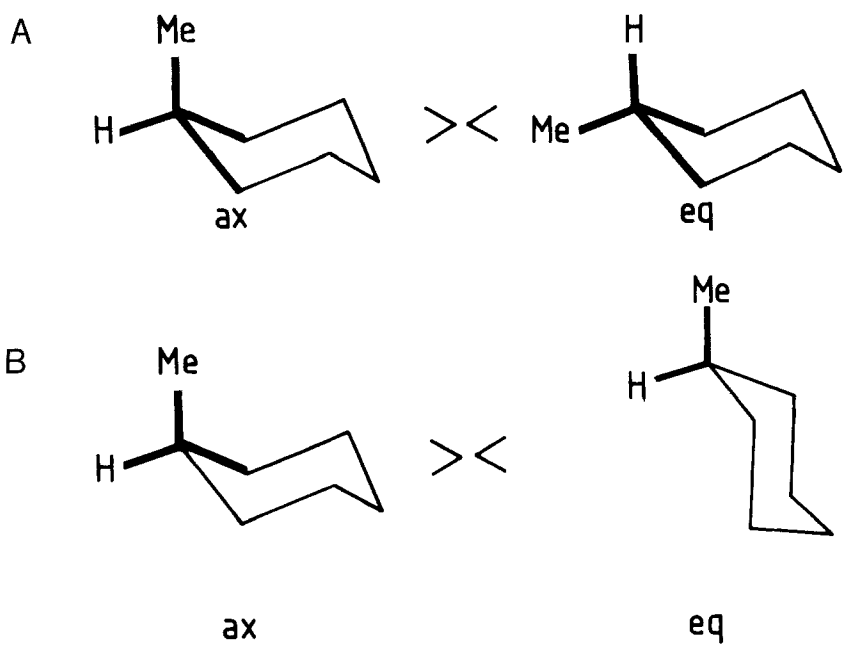


Fig. 2. Two of the ways to superpose the axial and the equatorial conformers of methylcyclohexane. The symbol calls for sliding the molecules onto each other until the thickened bonds coincide. A: Coincident rings; B: coincident H-C-Me fragments.

its molar mass (H-1 unit, C-12, etc.) and, for each molecule separately, a 3 × 3 matrix is constructed and diagonalised [13]. In the process, the origin shifts to the centre of mass (which coincides with the centre of symmetry, if there is one), axes and planes of symmetry shift to coincide with coordinate-axes and coordinate-planes, and as much mass as possible congregates about the origin, the x-axis and the xy-plane. The latter outcome has been exploited in the computerised drawing of molecules [14,15]. It amounts to retrieval of the cluster axis or plane, if any, and to the proper positioning of the cluster (straddling the x-axis or the xy-plane).

Obviously, all this does not ensure that the compromise is best as regards the coincidence of atomic and molecular elements of similarity. Still, if the juxtaposed molecules have an identical constituting unit, and if this unit comports a substantial portion of the whole, its sheer mass will force each molecule separately to orientate such that the outcoming mutual overlap be high. The same goes for weighty units that are not quite identical. In the example of methylcyclohexanes (Fig. 2), each isomer separately relaxes such that the symmetry plane ties to plane xz, the long axis C1...C4 approaches close to axis x, and the cluster-plane approaches plane xy (as in Fig. 2A, not 2B). The rings almost coincide, except for a small tilt-angle. This reduces somewhat the major overlap (between rings) but simultaneously enhances somewhat the minor overlap (between methyl groups).

It is worthwhile to generalise on this result. Stereochemical isomers differ in the orientation of a small portion of the molecule with respect to the main portion, but not in the chemical nature of either. The small portion does affect the course of geometrical transformation, but only to a small extent, and the two isomers are affected to roughly identical extents but in opposite senses. The outcome is that the major portions of the two isomers overlap almost entirely, except for a certain tilt that enables some overlap between the minor portions. We cannot stipulate that the resulting superposition is the best, i.e., that it corresponds to the maximal S. But we expect it to be close enough to the best, since the major portions overlap so well.

In the case of heteromolecules, these differ from hydrocarbons in comprising main atoms of variegated types. Molar masses spread over a range and atoms exert different extents of drive when transformed by mass. When molecules are compared, it is not necessary for atomic and molecular elements of similarity to attain good coincidence. As an example, the mass transformed coordinates of the main atoms in $CH_3Br$ are (x, y, z; nm)

| | | | |
|---|---|---|---|
| C | −0.163 | 0 | 0 |
| Br | 0.032 | 0 | 0 |

and those of $CH_3I$ are

| | | | |
|---|---|---|---|
| C | −0.191 | 0 | 0 |
| I | 0.024 | 0 | 0 |

Thus, the carbon atoms, which constitute an atomic element of similarity, are not coincident; the same goes by necessity for the hydrogen atoms and for the molecular elements defined by the methyl groups. The non-coincidence is not compensated by a beneficial overlap elsewhere: it is noteworthy that the shorter C-Br bond in $CH_3Br$ is not contained within the longer C-I bond in $CH_3I$.

TABLE 1
WEIGHTING FACTORS FOR SOME ATOMIC TYPES

| Atom | Mass (g/mol) | $r_w$ (nm) | $r_w^3$ ($nm^3 \times 10^3$) |
|---|---|---|---|
| H | 1 | 0.117 | 1.60 |
| C (tetra-covalent) | 12 | 0.175 | 5.36 |
| N- | 14 | 0.155 | 3.72 |
| -O- | 16 | 0.140 | 2.74 |
| O= | 16 | 0.150 | 3.38 |
| -S- | 32 | 0.180 | 5.83 |
| F | 19 | 0.130 | 2.20 |
| Cl | 35.5 | 0.177 | 5.55 |
| Br | 79.9 | 0.195 | 7.41 |
| I | 126.9 | 0.210 | 9.26 |

The situation is improved by using weighting factors that make main atoms more similar to one another. One conceivable choice is to transform 'by atomic volume', e.g., weight by some quantity that is proportional to a measure of atomic volume. The cubed van der Waals radius $r_w^3$ is such, and data for some atoms are assembled in Table 1. It is seen that atoms are frequently more similar to each other by the criterion of volume than by the criterion of mass. Since the requirements of non-identical atoms are now more similar than before, they position themselves more symmetrically with respect to elements of the coordinate system. In comparing $CH_3Br$ with $CH_3I$, one finds that Br in the former covers I in the latter:

| | | | | |
|---|---|---|---|---|
| $CH_3Br$, | C | −0.072 | 0 | 0 |
| | Br | 0.123 | 0 | 0 |
| $CH_3I$, | C | −0.093 | 0 | 0 |
| | I | 0.122 | 0 | 0 |

This, evidently, is but a partial step. At the extreme, one can attribute identical weights to all main atoms. We call this 'a transformation by biased mass', and assign weight 1 to H and weight 20 to all other atoms. In the example, this gives

| | | | | |
|---|---|---|---|---|
| $CH_3Br$, | C | −0.088 | 0 | 0 |
| | Br | 0.107 | 0 | 0 |
| $CH_3I$, | C | −0.097 | 0 | 0 |
| | I | 0.118 | 0 | 0 |

Now, the shorter C-Br bond is contained almost symmetrically within the longer C-I bond, with a consequent rise (admittedly small in this rudimentary example) in the similarity index (Eq. 4). S is 0.91 by mass, 0.92 by volume, and 0.93 by biased mass.

What we do in practice is calculate S four times – once for untransformed coordinates, thrice by the three transformations. The highest among the four numbers is the one we settle on. This is

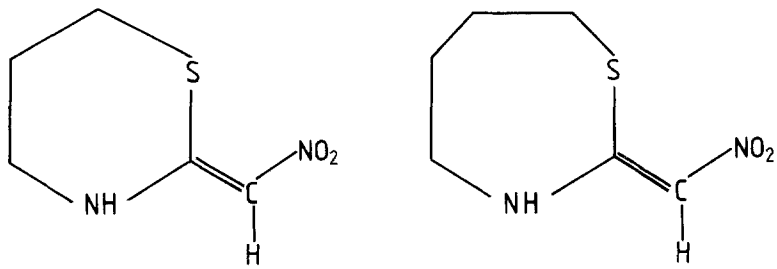


Fig. 3. Two similar biologically active nitroenamines.

justified, since the declared aim of the computation is to lay hands on the highest attainable value of S. Needless to add, the three transformations produce quite frequently close or identical orientations. When this is foreseeable, computation time may be saved by limiting the calculation to one sole transformation. For two molecules, each consisting of some 30 atoms, a run (transformation by one of the criteria plus point-counting at g = 0.02 nm and workup) requires about 7 cpu seconds on the VAX 11/780.

*Extent of movement*

In the geometrical transformation, it may be useful to have some measure of the extent of shift from the original orientation, in particular when the original orientation is not arbitrary or accidental. We obtain the measure by taking note of the shifts of molecular centres from their original positions. Let the shifts be $\Delta x_1$, $\Delta y_1$ and $\Delta z_1$ for molecule 1, and $\Delta x_2$, $\Delta y_2$ and $\Delta z_2$ for molecule 2. The relative shift R is then

$$R = [(\Delta x_1 - \Delta x_2)^2 + (\Delta y_1 - \Delta y_2)^2 + (\Delta z_1 - \Delta z_2)^2]^{1/2}$$

In our preliminary calculations, point-count similarities were compared with similarities based on quantum-chemical charge clouds [11]. One of the many cases examined was the pair of nitroenamine insecticides in Fig. 3 (by the notation of Ref. 11, these are 'Compd. 185', left, and 'Compd. 300', right). Originally, the molecules were placed with their fragments $C=CHNO_2$ coincident, which corresponded to S = 0.81. Results for the three types of transformation are as follows:

| | R (nm) | S | $S_H$ |
|---|---|---|---|
| Original | 0 | 0.812 | 0.811 |
| By mass | 0.0312 | 0.795 | 0.794 |
| By biased mass | 0.0386 | 0.790 | 0.788 |
| By volume | 0.0468 | 0.781 | 0.780 |

In this case (and, in fact, for the other nitroenamine pairs in the series [11]), freedom to shift lowers the calculated S, but the lowering is virtually insignificant: if $C=CHNO_2$ superposition were not tried, and the program was allowed to choose its own preferred orientation, S would

have come out equal to 0.79 (by mass) rather than to 0.81. The quantum-chemical similarity amounts to 0.80 (by electric field), with $C=CHNO_2$ moieties shifted somewhat from perfect coincidence.

Clearly, the results just quoted are not necessarily typical, but results of this type are not rare. This is because there is more reason to calculate the similarity of similar molecules than that of dissimilar molecules, and the similarity of similar molecules is based largely on the presence of some common fragment. In calculating starting geometries for similar molecules, it is natural to start by building up the common fragment, adding to it later and separately the other atoms in each of the two molecules.

*Use as a measure of chirality*

The notion of chirality is couched in terms of symmetry elements, and therefore chirality is a 'black-or-white' property: a molecule is either chiral or achiral. It has been argued, though, that such an attitude risks overlooking fine details of molecular shape, and that it would be useful to quantify the intuitive feeling that, say, 2-deuteriobutane diverges less from the achiral form than does 2-bromobutane [4]. Likewise, it would be helpful to have a measure of chirality that correlates with the experimental finding that, say, the molecular rotation of MeCHClEt is larger than that of $MeCH(CH_2Cl)Et$. In a recent article [3], Gilat writes: '... It becomes possible to assign a chiral coefficient $\chi$ to any chiral molecule or a unit cell in a crystal. As it now stands, this may involve a considerable amount of mathematical and computational effort, but the definition is valid'.

His recipe runs as follows. Given a chiral body, its two enantiomers are to be superposed in a way to maximise the volume of overlap. The total volume V and the volume $V_o$ of overlap are to be calculated. When the enantiomers are superposed, each leaves a volume $V - V_o$ dangling outside the zone of overlap. The coefficient of chirality is then $\chi = (V - V_o)/V$.

Now, maximisation of overlap is the essence of similarity computations. The volumes V and $V_o$ are related to our T and B through a common proportionality constant k: $V = kT$ and $V_o = kB$ (cf. Eq. 3), where k is the volume associated with one point of grid, $k = g^3$. For enantiomers, then,

$$\chi = \frac{V-V_o}{V} = \frac{T-B}{T} = \frac{(O+B)-B}{O+B} = \frac{O}{O+B} = D$$

Thus, the coefficient $\chi$ of chirality is nothing more than our measure of dissimilarity, D (Eq. 7). Being such, its evaluation is actually devoid of any computational effort. For CHFClBr we calculate $S = 0.899$ and $D = 0.101$ (idealised tetrahedral bond angles, standard bond lengths); Gilat estimates $\chi = 0.088$ (details not given). Note that demonstration of the $\chi$/D identity is feasible only within the point-counting formulation, not with the original charge-cloud definitions of molecular similarity (Eqs. 1 and 2).

D-values for some enantiomeric pairs are listed in Table 2.

## DISCUSSION

Linear alkanes provide a good starting point. Let $C_mH_{2m+2}$ and $C_nH_{2n+2}$ be two such molecules in their stretched (zigzag) conformation. They are more similar to each other to the extent

TABLE 2
COMPUTED DISSIMILARITY OF ENANTIOMERS

| | | | D |
|---|---|---|---|
| 1. | $CHFClBr$ | | 0.10 |
| 2. | $CH_3CHClCH_2CH_3$ | Cl/Me-anti | 0.20 |
| | | Me/Me-anti | 0.23 |
| 3. | $CH_3CHBrCH_2CH_3$ | Br/Me-anti | 0.16 |
| | | Me/Me-anti | 0.28 |
| 4. | $CH_3CHICH_2CH_3$ | I/Me-anti | 0.13 |
| | | Me/Me-anti | 0.26 |
| 5. | $CH_3CH_2CH_2CH_3$ | gauche | 0.17 |
| 6. | $C_6H_{11}OH$ | HCOH, gauche | 0.18 |
| 7. | 1,2-Dimethylcyclohexane | ax, eq | 0.20 |
| | | ax, ax | 0.23 |
| | | eq. eq | 0.28 |
| 8. | $CH_3CHClCHClCH_3$ | dl, Cl/Cl-anti | 0.23 |
| 9. | $FCH_2CH_2F$ | gauche | 0.21 |
| 10. | $ClCH_2CH_2Cl$ | gauche | 0.34 |
| 11. | $BrCH_2CH_2Br$ | gauche | 0.40 |
| 12. | $ICH_2CH_2I$ | gauche | 0.42 |

that (a) n is closer to m, and (b) m and n are larger. Thus, decane is more similar to nonane than to octane, and decane is more similar to nonane than is pentane to butane. Computed similarities of decane to shorter homologs are as follows (S, and $S_H$ in parentheses): pentane, 0.64 (0.61); hexane, 0.77 (0.75); heptane, 0.85 (0.84); octane, 0.90 (0.89); nonane, 0.95 (0.95). The numbers illustrate the generalisation that $S > S_H$ (case 1 below Eqs. 4 and 5), and also show that the difference $S - S_H$ diminishes with the difference $n - m$. This is to be anticipated since, at the limit, $S = S_H$ for $n = m$ (case 2).

The numbers above provide convenient references to convey similarities among other molecules. This is often helpful, since intuition grasps the similarity of straight-chain alkanes better than that of more complex organic molecules. Thus, intuition can hardly visualise the information that $S = 0.81$ for the two molecules in Fig. 3. The number becomes, however, more meaningful when the molecules are described as 'more similar to each other than hexane is to decane, but not as similar as is heptane to decane'.

We have seen that S increases with B (Eqs. 4 and 5), and hence with kB, the size of the fragment that is common to the juxtaposed molecules. When a sequence of related pairs is examined, this opens some room for confusion. An example is provided by the diaxial and diequatorial varieties of *trans*-1,2-dichlorocyclohexane, $C_6H_{10}Cl_2$, and the corresponding 4-eq-*tert*-butyl derivatives, $(CH_3)_3C\text{-}C_6H_9Cl_2$. For the non-derivatised pair, $S = 0.76$ ($g = 0.02$ nm, by volume); subsequent to derivatisation, the similarity leaps to 0.84. This is not because the interesting dichlorocyclohexyl moieties have become more similar to each other. Rather, it is because the bulky $(CH_3)_3C$-unit has been appended to the isomers, thereby increasing B: it goes up from ca. 12 500 to ca. 20 800 points. In fact, $O = O_1 \sim O_2$ is about 3900 for both pairs.

It is interesting to grade the similarities of isomeric species. The two rotational isomers of

$C_6H_{11}OH$ (H rotates about C-O) have S = 0.97; the staggered and eclipsed rotamers of ethane and methanol are characterised, respectively, by S-values of 0.90 and 0.89. Next come axial-equatorial stereoisomers in substituted six-membered rings, like axial and equatorial methylcyclohexane, endo and exo 2-methylnorbornane (both 0.82), diaxial and diequatorial 1,2-dimethylcyclohexane or dichlorocyclohexane (both 0.76). In this range fall also the pairs of xylenes (ortho vs. meta, ortho vs. para, meta vs. para, all 0.82). Stereoisomers of chain-conformation or ring-configuration come next. Examples are anti and gauche 2,3-dimethylbutane (0.76), diaxial and axial-equatorial 1,2-dimethylcyclohexane (0.78).

Among isomers of position, greater similarity occurs in those pairs wherein the branches are identical, like 2,2-dimethylbutane and 2,3-dimethylbutane (anti, 0.74; gauche, 0.72), or 1,2-diequatorial and 1,4-diequatorial dimethylcyclohexane (0.70). Least similar are pairs where branches have different lengths, like 2,4-diethylpentane and 2,2,4,4-tetramethylpentane (0.65).

Two groups of heteromolecule prototypes that we examined systematically are the interhalogens X-Y and the halomethanes $CH_3$-X (X, Y = F, Cl, Br, I). Similarities among interhalogens are cited in Table 3. They cover a very wide range, 0.67–0.93 (by biased mass). The sequence is revealing as regards the interplay of atomic and molecular elements of similarity. As can be seen, similarity is affected simultaneously by the types of halogen and the interatomic distances [17]. The least similar pair in Table 3 is ClF/IBr: not only are the two molecules made of totally different atoms, where ClF contains the lightest and IBr the heaviest halogens, but also bond Cl-F is the shortest among interhalogens (0.1628 nm) while I-Br is the longest (0.2485 nm). The most similar pairs in the Table, BrCl/ICl and ICl/IBr, are composed of three molecules such that (a) each pair has one atom in common, and (b) bonds are close in length and consecutive in the list of bond lengths [16].

TABLE 3
SIMILARITIES AMONG INTERHALOGENS

| S | Pair | Δl (nm) |
|---|---|---|
| 0.93 | BrCl/ICl | 0.0183 |
| | ICl/IBr | 0.0164 |
| 0.91 | BrF/IF | 0.0153 |
| 0.89 | ClF/BrF | 0.0128 |
| 0.88 | BrCl/IBr | 0.0347 |
| 0.86 | IF/ICl | 0.0412 |
| 0.84 | IF/BrCl[a] | 0.0229 |
| | BrF/BrCl | 0.0382 |
| 0.81 | ClF/IF | 0.0281 |
| 0.80 | BrF/ICl[a] | 0.0565 |
| | IF/IBr | 0.0576 |
| 0.77 | ClF/BrCl | 0.0510 |
| 0.75 | BrF/IBr | 0.0729 |
| 0.71 | ClF/ICl | 0.0693 |
| 0.67 | ClF/IBr[a] | 0.0857 |

[a]Molecules with alien atoms.

To emphasise this point, Table 3 lists also the difference Δl in bond length between members of each pair. Roughly, molecules with small Δl are more similar than molecules with large Δl.

The sequence of similarities among halomethanes follows a similar pattern. Calculations, based on standard geometries with tetrahedral angles [17], led to the following series: $CH_3Br/CH_3I$ ($\Delta l = 0.0200$ nm, most similar, $S = 0.93$) > $CH_3Cl/CH_3Br$ (0.0154) > $CH_3Cl/CH_3I$ (0.0354) > $CH_3F/CH_3Cl$ (0.0403) > $CH_3F/CH_3Br$ (0.0557) > $CH_3F/CH_3I$ (0.0757, least similar, $S = 0.70$). In this series, the dependence on bond length is marked even more than in the interhalogens.

## CONCLUSION

We have presented a simple recipe to approximate the indices S and $S_H$ of molecular similarity. Input requires no more than the atomic coordinates of two molecules, and computation consists of the diagonalisation of a few 3 × 3 matrices. We have cited some numerical results and made a few generalisations as regards similarity elements and gradation in the magnitude of the indices.

One may ask at what S do two molecules become sufficiently similar, e.g., for the ends of drug design. A scan of similarities among nitroenamine insecticides (of which the pair in Fig. 3 is one instance) suggests that the threshold lies at about $S = 0.80$. There is a caveat, though. As discussed above, substitution by a bulky residue at an unimportant zone of the similar molecules raises the computed similarity even if the region of interest is unaffected. The reason is that B increases appreciably while $O_1$ and $O_2$ are not changed by much (Eqs. 4 and 5). By this token, if the vinyl hydrogen of the molecules in Fig. 3 were replaced in both by some bulky group, S would rise even though skeletal resemblance is not increased. In fact, such a substitution could affect adversely the resemblance between the relaxed ring-conformations in the two molecules, while the calculated similarity would rise nevertheless.

One may also ask which type of electron-density similarities (Eqs. 1 and 2) do our shape-similarities (Eqs. 3 and 4) mimic: all-electron or valence-electron values. Comparison with available data [10] suggests that ours are close to valence-electron numbers, which is considered to be an advantage. The data in question are the similarities among propane ($C_3H_8$), dimethyl ether ($Me_2O$) and dimethyl sulphide ($Me_2S$). From ab initio molecular orbitals one derives the following S-values. All electrons: $C_3H_8/Me_2O$, 0.67; $C_3H_8/Me_2S$, 0.36; $Me_2O/Me_2S$, 0.50. Valence electrons only: 0.87, 0.95 and 0.75, in respective order. Our numbers are 0.91, 0.95 and 0.88, i.e., closer to the latter than to the former series.

Note that drug design is only one possible field of application of similarity indices [2]. Studies related to the Hammond postulate [18], where similarities of the activated complex to the reactant and to the product of reaction are to be compared, constitute another. So do attempts to devise stable structures which mimic transition states, as does penicillin. A more academic usage could be the comparison of two different structures, measured for a given molecule or calculated by two different techniques, or of one measured and one calculated. Thus, on comparing the microwave with the electron-diffraction geometry of dimethyl sulphide [19], we get $S = 0.99$, that is, the dissimilarity in this particular case does not exceed 1 %.

Finally, a measure of molecular dissimilarity has an interest of its own, in particular in the context of enantiomeric pairs. We argue above that there is sense in grading the similarity of chiral molecules (and chiral objects in general) to their enantiomers. The enantiomers of $CH_3CHDCH_2CH_3$, for example, resemble each other more than do the enantiomers of

$CH_3CHBrCH_2CH_3$, since D resembles H more than does Br, and replacement of D or Br by H makes all species in question identical and achiral. In other words, $CH_3CHDCH_2CH_3$ is closer than $CH_3CHBrCH_2CH_3$ to the parent achiral form. Now, the notion of chirality is based on the absence of a certain type of symmetry elements and, as such, is not quantifiable: the elements are either there or not there. Similarity elements, by contrast, do allow for variation, since the similarity of atoms, bonds and angles does not imply absolute identity. Hence, quantitative ordering of chiral molecules can be derived from a measure of dissimilarity. The obvious definition of this measure ($D = 1 - S$) not only turns out to be equivalent to a previously proposed 'coefficient of chirality', but also renders straightforward and simple the evaluation of that coefficient.

## ACKNOWLEDGEMENT

This work was conducted in part pursuant to a contract with the National Foundation for Cancer Research.

## REFERENCES

1 Carbo, R., Leyda, L. and Arnau, M., Int. J. Quantum Chem., 17 (1980) 1185.
2 Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem., Quantum Biol. Symp., 14 (1987) 105.
3 Gilat, G., J. Phys. A: Math. Gen., 1989, L545 (1989).
4 Avnir, D. and Meyer, A.Y., J. Mol. Struct. (Theochem.), 179 (1988) 83.
5 a. Meyer, A.Y., Chem. Soc. Rev., 15 (1986) 449.
b. Meyer, A.Y., Struct. Chem., 1 (1990) 265.
6 Gavezzotti, A., J. Am. Chem. Soc., 105 (1983) 5220, and 107 (1985) 962.
7 Meyer, A.Y., J. Comput. Chem., 9 (1988) 18, and previous papers in the series.
8 Allinger, N.L., J. Am. Chem. Soc., 99 (1977) 8127.
9 Meyer, A.Y., J. Mol. Struct., 94 (1983) 95, and previous papers in the series, and 105 (1983) 143.
10 Hodgkin, E.E., Molecular Similarity in Computer-Aided Molecular Design, Doctoral Thesis, University of Oxford, 1987.
11 Burt, C., Richards, W.G. and Huxley, P., J. Comput. Chem., 11 (1990) 1139.
12 Lister, D.G., Macdonald, J.N. and Owen, N.L., Internal Rotations and Inversions, Academic Press, London, 1978.
13 Hirschfelder, J.O., J. Chem. Phys., 8 (1940) 431.
14 Shmueli, U., J. Mol. Graphics, 2 (1984) 111.
15 Meyer, A.Y., J. Comput. Chem., 7 (1986) 144.
16 For a compilation of interhalogen properties see: Bailar, J.C., Emeléleus, H.J. and Nyholm, R. (Eds.) Comprehensive Inorganic Chemistry, Vol. 2, Pergamon Press, Oxford, 1973, Table 93 (pp. 1487–1500). Bond lengths are (re, nm): ClF, 0.1628; BrF, 0.1756; IF, 0.1909; BrCl, 0.2138; ICl, 0.2321; IBr, 0.2485.
17 The bond lengths used are (nm): C-H, 0.1113; C-F, 0.1392; C-Cl, 0.1795; C-Br, 0.1949; C-I, 0.2149.
18 Hammond, G.S., J. Am. Chem. Soc., 77 (1955) 334.
19 a. Pierce, L. and Hayashi, M., J. Chem. Phys., 35 (1961) 479.
b. Iijima, T., Tsuchiya, S. and Kimura, M., Bull. Chem. Soc. Jpn., 50 (1977) 2564.