The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity

Lowell H. Hall*

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Lemont B. Kier†

Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond, Virginia, 23298

Received November 7, 1999

The electrotopological state (E-state) is presented as a representation of molecular structure useful for definition of a space for chemical structures. This E-state representation provides the basis for chemical database management. The E-state formalism is presented along with its extension to the atom-type E-state. An approach to database organization, using polychlorobiphenyls (PCBs) as examples, reveals the descriptive power of the E-state paradigm. A well-organized chemical database, as described here, may be searched to find structures similar to a target structure with the expectation that such structures may exhibit properties similar to the target. Searches using the atom-type E-state indices are demonstrated with two example drug molecules.

INTRODUCTION

In any process of molecular modeling, the need for a representation of molecular structure is critical and its role is significant. An information-rich representation which is rapidly computed and readily manipulated is essential. Ease of interpretation is also necessary, especially when interpretation entails well-known chemical intuition. The electrotopological state (E-state) formalism and resulting descriptors possess all these characteristics, as demonstrated in a recent book[1], a more recent book chapter,[2] and many papers.[3−9]

Atom Information Fields. Any approach to the quantitative description of atoms within a molecule must be built upon both their context and their relationships which operate within the molecule. We can view each atom in a molecule as existing in a field within that molecule in which all other atoms participate, a field which we have termed an information field.[1,4] Each atom possesses intrinsic qualities which are modified by all the other atoms within the molecule. The methyl group in propanoic acid is different from the methyl group in methylamine by virtue of their different chemical environments. The two methyl groups in *N*-methylpropanamide are also different from each other. Quantifying the methyl group requires both a description as a methyl group and its relationship to all other atoms in the molecule in which it resides. The influence of all other atoms in propanoic acid or methylamine makes the methyl group unique relative to methyl groups in all other molecules. The intrinsic characteristics of the methyl group which transcend its context must be identified in order to describe it in a quantitative way. We consider each of these characteristics in a brief reprise of the E-state formalism.

The Intrinsic State of an Atom. In propanoic acid we recognize that the methyl group possesses attributes that are identified as *intrinsic* and also distinguish it from the adjacent methylene group even though both contain C sp^3 atoms. The common attributes of methyl groups that comprise its intrinsic state are those which we believe have an important influence on chemical, physical, and biological properties. For any atom (or hydride group such as $-CH_3$, $-NH_2$, and $-OH$, also often referred to as atoms) we can identify three attributes clearly.

The *first* attribute is elemental content. The methyl is composed of carbon and hydrogen atoms. A single atom in a molecule is simply described as that element.

The *second* attribute of the intrinsic state is electronic organization, which can be represented by the hybrid state or by the valence state of the atom. This description includes counts of σ, π, and lone pair electrons comprising the valence electrons of the atom. In the case of a group with bonded hydrogen atoms, we also encode the number of hydrogens to distinguish it from the other hydrides of carbon.

A *third* attribute in an intrinsic state description is the degree of adjacency or more generally the topological state of the atom (hydride group). This attribute is important in defining the position of the atom relative to the topology of the whole molecule. As an example, if we compare carbon atoms in the three isomers of pentane, it is apparent that the spatial accessibilities of some carbons are different. The methyl groups in pentane reside on the periphery of the molecule (and are sometimes called mantle fragments); hence, they are easily accessible to interactions with neighboring molecules. In contrast, the methylene group in pentane is located within the structure of the molecule with somewhat diminished accessibility to an intermolecular contact. The branched atom of isopentane is partially surrounded within the molecule. Its accessibility to any intermolecular contact

* To whom correspondence should be addressed. Telephone: 617 745 3550. E-mail: hall@enc.edu.

† Telephone: 804 828 6451. E-mail: kier@gems.vcu.edu.

10.1021/ci990140w CCC: $19.00 © 2000 American Chemical Society
Published on Web 04/04/2000

is much smaller than that of the methyl or methylene group. Finally, the quaternary carbon at the center of the neopentane molecule is buried, and its intermolecular accessibility is virtually nil.

In the following presentation, we demonstrate a way to quantify these three attributes.

The Graph Representation of a Molecule. A common approach to the description of a molecule is the use of the chemical graph. The bonding scheme of atoms is depicted as a network using lines to represent connections between atoms (hydride groups) and vertices to represent atoms (hydride groups). It is common to omit hydrogen atoms from carbon vertices since these are implied by valence rules. Hydrogen atoms associated with heteroatoms are retained in the scheme for clarity. Multiple bonds are explicitly represented. Aromatic rings include π bonds using canonical representation or a circle denoting the π orbital annulus. We begin our description of a molecule by writing the conventional chemical structure of a molecule such as *N*-methylpropanamide, Figure 1, and then delete hydrogens associated with carbon atoms to form the chemical graph, often called the hydrogen-suppressed graph.

The δ Values. To encode information on electron counts, we determine for each atom two values previously defined in molecular connectivity descriptions.[10] First we consider the simple delta value, δ, which is the count of adjacent atoms (other than hydrogen) in the molecular skeleton. This definition is equivalent to describing just the σ bond skeleton network. The simple δ value for an atom is given as $\delta = \sigma - h$. The second delta value is calculated for each atom based on the total number of valence electrons on that atom minus the number of bonding hydrogens. This quantity is called the valence δ values, δ^v: $\delta^v = \sigma + \pi + n - h$.[10,11]

These two sets of δ values, illustrated in Figure 1b, are the basic ingredients for the definition of the intrinsic state of atoms in molecules.

Electronic Information in the δ Values. The pair of δ values just discussed provide a rich resource of information which describes atoms in molecules in our quest for a structural description. It is summarized as follows:

The simple delta value, δ, encodes (a) the count of adjacent atoms excluding hydrogen, (b) the count of σ electrons on an atom excluding hydrogen, (c) the count of bonds joining an atom to other than hydrogen, and (d) the topological environment of the atom in the molecule.

The valence delta value, δ^v, encodes (a) the count of valence electrons on an atom other than to hydrogen and (b) the count of σ, π, and lone pair electrons excluding bonds to hydrogen.

Table 1 summarizes the δ and δ^v values for carbon, nitrogen, oxygen, and fluorine atoms in a molecule. It is evident that $\delta^v - \delta = \pi + n$, the count of π and lone pair electrons on an atom in a molecule. This information provides a quantitative measure of the potential of the atom for intermolecular interaction and reaction. These values comprise the Kier–Hall electronegativity[11] which has a high correlation ($r = 0.988$) with the Mulliken–Jaffe electronegativity of atoms in their valence states,[12] shown in Table 1. The electronegativity of an atom in a molecule is of major importance within the context of the general information field described earlier. As a consequence this simple statement of structure has a significant role in encoding the intrinsic state of the atom.

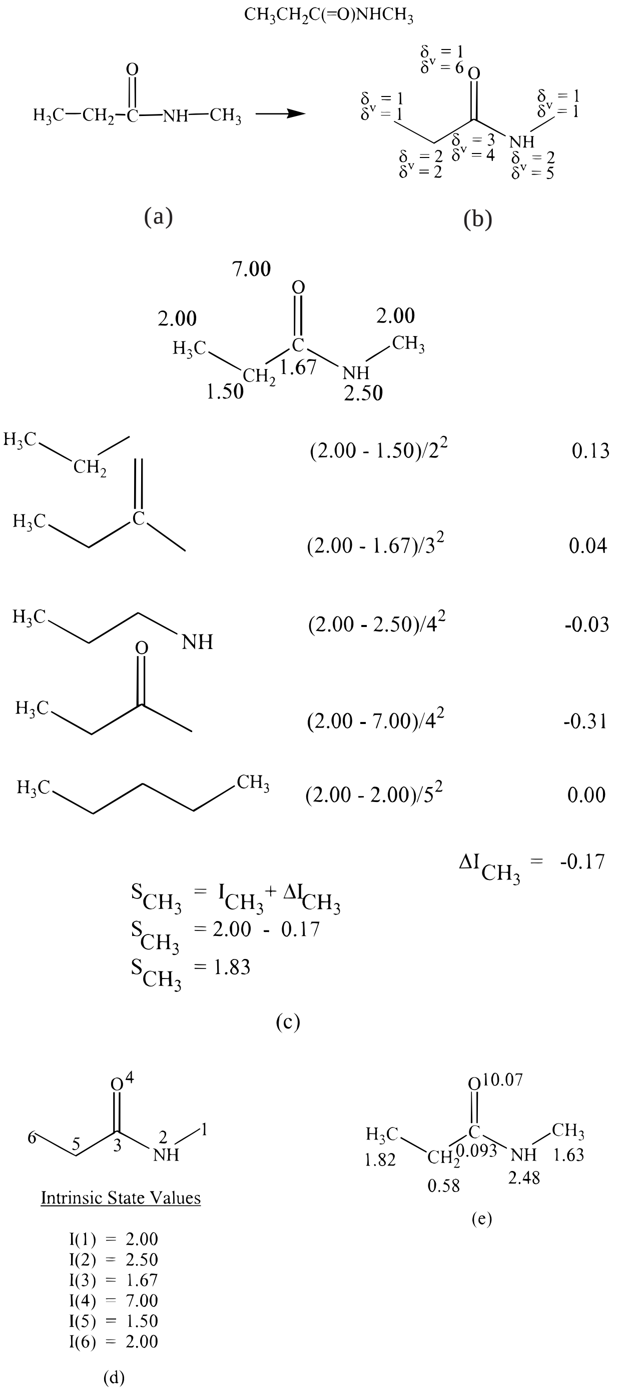


Figure 1. (a) Molecule of *N*-methylpropanamide. (b) *N*-Methylpropanamide presented as a chemical graph with δ and δ^v values. (c) *N*-Methylpropanamide showing the computation of *S* for one methyl group. (d) *N*-Methylpropanamide with calculated intrinsic state values, *I*. (e) *N*-Methylpropanamide with calculated E-state values.

Intrinsic State Algorithm. The possibility of encoding a close approximation of the valence state electronegativity with such a simple index is of great value. Table 1 lists the values of δ^v and δ for several covalently bound atoms in their valence states. The derivation of an intrinsic state index labeled *I*, begins with the use of the $\delta^v - \delta$ term. Of equal importance in defining an intrinsic state is the adjacency or topology of the atom in the molecule. Accordingly, the intrinsic state encodes two attributes: (1) the availability of

Table 1. δ Values and Kier−Hall Electronegativities of Atoms

atom	δ	δ^v	X_{KH}[a]	X_M (eV)
−F	1	7	1.50	17.36
=O	1	6	1.25	17.07
−O−	2	6	1.00	15.25
≡N	1	5	1.00	15.68
=N−	2	5	0.75	12.87
>N−	3	5	0.50	11.54
≡C−	2	4	0.50	10.39
=C<	3	4	0.25	8.79
>C<	4	4	0.00	7.98
−Cl	1	7	0.67	11.54
=S	1	6	0.55	10.88
−S−	2	6	0.44	10.14

[a] $X_{KH} = (\delta^v - \delta)/n^2$: n = principal quantum number; X_{KH} = Kier−Hall electronegativities; X_M = Mulliken−Jaffe electronegativities expressed in electronvolts.

Table 2. Intrinsic State Values of Second Row Hydrides

atom hydride group	$I = [(\delta^v + 1)/\delta]$	atom hydride group	$I = [(\delta^v + 1)/\delta]$
>C<	1.250	>C<	1.250
>CH−	1.333	>CH−	1.333
$-CH_2-$	1.500	$-CH_2-$	1.500
>C=	1.667	>C=	1.667
$-CH_3$, =CH−, >N−	2.000	$-CH_3$, =CH−, >N−	2.000
≡C−, −NH−	2.500	≡C−, −NH−	2.500
$=CH_2$, =N−	3.000	$=CH_2$, =N−	3.000
−O−	3.500	−O−	3.500
≡CH, $-NH_2$	4.000	≡CH, $-NH_2$	4.000
=NH	5.000	=NH	5.000
≡N, −OH	6.000	≡N, −OH	6.000
=O	7.000	=O	7.000
−F	8.000	−F	8.000

the atom or group for intermolecular interaction and (2) the manifold of bonds over which adjacent atoms may influence and be influenced by its state.

The adjacency, encoded by the simple delta value, δ, must therefore be a companion descriptor with the electronegativity in defining the intrinsic state. One approach is to take the ratio of the valence state electronegativity to the degree of adjacency as the intrinsic state. In this manner the greater degree of adjacency, δ, the less the accessibility; the smaller this value, the greater the accessibility of an atom or group. The ratio of the two terms produces an initial description of the intrinsic state value, I. (This expression was processed into a definition shown later.)

$$I\colon\ (\delta^v - \delta)/\delta \tag{1}$$

In this form, the intrinsic state, I, may be viewed as the ratio of the π and lone pair electron count to the count of avenues of intramolecular interaction, the number of σ bonds in the skeleton for this atom. That is, due to the intramolecular interaction associated with the atom, π and lone pair electron density may be influential across the bonding network which is the set of σ bonds in the molecular skeleton.

The value found for I in eq 1 for the carbon sp^3 atom is zero since $\delta^v = \delta$ in each case (e.g. $-CH_3$, $-CH_2-$, >CH−). If we scale the $\delta^v - \delta$ term by adding 1, the zero values are eliminated and there is a discrimination among the various hydrides of carbon, arising from the different values of δ. This modification leads to eq 2. This expression

$$I\colon\ (\delta^v - \delta + 1)/\delta \tag{2}$$

achieves the objective of encoding electronic structure and topology through an approximation of the valence state electronegativity and local neighborhood of bonds. An alternative form of this expression reveals the π and lone pair count explicitly: $(\pi + n + 1)/(\sigma - h)$.

A simplification of eq 2 can be made by scaling the expression, that is, by adding 1 to the entire term to produce an expression for the intrinsic state for an atom or hydride group in a molecule (in the second row of the periodic table).

$$I\colon\ (\delta^v + 1)/\delta \tag{3}$$

Table 2 shows the intrinsic states of second row atoms and groups. Figure 1d shows intrinsic state values for N-methylpropanamide. To calculate the I values for higher quantum level atoms, the valence δ value is calculated as $(2/N)^2\delta^v$, where N is the principal quantum number, giving the formal definition of the intrinsic state:

$$I \equiv ((2/N)^2\delta^v + 1)/\delta \tag{4}$$

Field Influences on the Intrinsic State. We have derived the intrinsic state of an atom or group. This expression does not reflect its effect on other atoms or its influence within the field of other atoms in a molecule. This reciprocal influence may take the form of a perturbation of the intrinsic state using some characteristic of every other atom in the molecule. A reasonable choice of attributes producing a perturbation on an atom is the intrinsic states of all other atoms in the molecule. This approach is analogous to using electronegativities of other atoms to modify the electronic state of each atom within the field of the molecular structure. The vehicle for this influence is the network of bonds linking each atom with all others in the molecule. This network is synonymous with the chemical graph model of the molecule over which electronegativity influence manifests itself.

A second consideration is the influence of two atoms in a molecule on the intrinsic state of the other. Since the chemical graph is the model of the presence and connectivity of atoms within the molecule, the count of bonds or atoms in paths separating two atoms was chosen for the unit of distance between any two atoms in a molecule. More precisely, the count of atoms in the minimum path length, r_{ij}, separating two atoms, i and j, is the distance selected to encode the influence between two atoms. Note that this count is equal to the usual graph distance plus 1. From this model we have chosen the difference between the intrinsic states of atom i and atom j, $(I_i - I_j)$, as the perturbation on each other. This effect is assumed to diminish by a power, m, of the distance; hence, the perturbation of I_i, called ΔI_i, is expressed as

$$\Delta I_i = (I_i - I_j)/{r_{ij}}^m \tag{5}$$

The choice of the value of m results in a variable influence of distant and close atoms in the graph. Most studies to date have employed a value of $m = 2$.[1] The total perturbation of atom i is a consequence of the influence of all other atoms in the molecule. Accordingly, the total perturbation of the intrinsic state of atom i, ΔI_i, should be a sum of these individual perturbations. The term, $\Sigma\Delta I_{ij}$, is a sum of all perturbation terms expressed by eq 5. Figure 1c shows all the terms affecting one of the methyl groups in *N*-methylpropanamide. The state of atom *i* in a molecule is the intrinsic state, I_i, plus the sum of all perturbations included in ΔI_{ij}. The result for atom *i* is called the electrotopological state, S_i, and is expressed as

$$S_i = I_i + \sum_j \Delta I_{ij} \tag{6}$$

For brevity, the S_i term is called the E-state for atom *i*. Figure 1e shows the E-state values calculated for *N*-methylpropanamide. The E-state index was introduced in a series of articles[3−9] and is fully described with many examples in a recent book.[1]

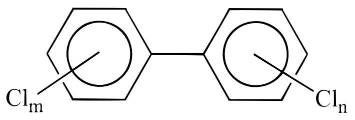
(a) Atom-Type E-State Indices. An extension of the E-state formalism is the use of an atom-type index making it possible to study molecules of noncongeneric structure. Each atom is classified according to its valence state, the number of bonded hydrogens, and aromaticity.[13] For an atom-type index, the E-state values are summed for all atoms of the same type in the molecule. The symbol for an atom-type index is $S^{T}(X)$, where X denotes the atom or hydride group. As examples we have $S^{T}(-\text{Cl})$ for a chloro, $S^{T}(-\text{OH})$ for a hydroxyl group, and $S^{T}(\overline{\cdots}\text{CH}\overline{\cdots})$ for an aromatic CH. The software used to compute all indices is Molconn-Z which recognizes 80 atom-types.[14]

Atom-type indices encode three distinct types of chemical structure information: (1) electron accessibility for the atoms of that type, (2) the presence/absence of the atom-type, and (3) a count of atom-type present in the molecule.

The atom-type E-state indices are used for heterogeneous data sets for structure−activity and for similarity analyses.[1]

(b) Organization of Databases. Organization of chemical databases has been discussed by many authors, including two seminal books.[15,16] In this paper we are presenting the use of the E-state indices as the basis for representation of the structures in chemical databases. The atom-type E-state values for atoms or groups in a molecule may be considered as basis vectors in a structure space, that is, a manifold containing all possible atoms or groups. Each dimension is an atom-type E-state calculated for a particular type of atom or group, that is, the sum of the individual E-state values for all atoms (groups) of a given type. In general, the atom-type E-state indices are not highly intercorrelated. They form a very nearly orthogonal basis set of descriptors largely because of the independence of the presence and count of various groups in a database.

Within a subset of molecules in a database there are patterns of structure that are of interest in compound design. A pattern is a sequence of structures which varies in some systematic and describable manner. These patterns characterize the relative similarity or diversity within a subset of structures and between sets. The atom-type indices make it possible to organize the subset in some manner, which facilitates the design of other modifications and the selection of diverse structures for testing or cluster analysis based on structure, and to conduct structure−activity analyses. In E-state space one may consider the distance between structures, characterize clusters, and examine patterns of structures in the neighborhood of interest. In the next section we illustrate this organization and show how the E-state indices can accomplish this by examining the notorious polychlorobiphenyls (the PCBs) (I).



I

DISCUSSION

Polychlorobiphenyls. Three atom-types are present in this series, each represented by its atom-type E-state code. These atom-types are the chlorine atom-type, $S^{T}(-\text{Cl})$, the aromatic CH group, $S^{T}(\overline{\cdots}\text{CH}\overline{\cdots})$, and the substituted aromatic carbon atom,

$$S^{T}(\overline{\cdots}\overset{|}{\text{C}}\overline{\cdots})$$

Using the sum of E-states for each atom-type as described above, a structure space is created. To illustrate the structure organization, the two parameters, $S^{T}(-\text{Cl})$ and $S^{T}(\overline{\cdots}\text{CH}\overline{\cdots})$, are used to create a two-dimensional space which is a projection from the larger space of all atom types. A view of this space over a relatively wide range reveals many of the possible polychlorobiphenyls from the unsubstituted through the tetrasubstituted derivatives, Figure 2. The major parameter governing the position in this space in a coarse manner is the count of the number of chlorine atoms. It is notable that a group of compounds such as the trichloro-PCBs are spread in a cluster rather than superimposed on a single point, as they would be if atom counts were used to define the space. To extract more useful information, we focus on a subset with a common number of chlorine atoms by viewing a restricted range of parameter values.

In Figure 3 are shown the dichlorobiphenyls in the atom-type E-state space for Cl and aromatic C−H coded as $S^{T}(-\text{Cl})$ and $S^{T}(\overline{\cdots}\text{CH}\overline{\cdots})$. This figure presents an opportunity to assess the structure information as organized by these two parameters. The disubstituted biphenyls are arrayed in a pattern that can be interpreted directly in terms of structure. The structures are shown in Figure 3 and identified in three boxes which are characterized by specific structure characteristics. The upper box contains the only two molecules, (2,2′) and (2,6), which are substituted on the rings at the 2 or 2′ positions in (*I*), that is, ortho to the bond which adjoins the two phenyl rings (hereafter called simply ortho positions). The lowest box contains molecules with chlorine atoms distant from the juncture of the two rings (meta and para positions). Between these two subsets lie biphenyls substituted at intermediate positions (meta or para and one ortho). At the top of the middle box one chlorine is in the ortho position while the other is in a meta position. At the bottom of the box, one chlorine is ortho and the other is in the para position. Likewise in the bottom box, at the top are isomers with two meta chlorines; at the bottom is the only 4,4′ isomer. There is a clear trend in the structures in the directions from the top to the bottom of Figure 3.

In addition to the ortho, meta, para characterization, a second pattern may also be described. PCBs to the right of the dashed line have both chorines on the same ring, whereas on the left the two chlorines are on separate rings. As one moves diagonally from top left to bottom right, the chlorines go from close together at the ortho positions to far apart at the para positions. This set of PCBs is clearly organized in

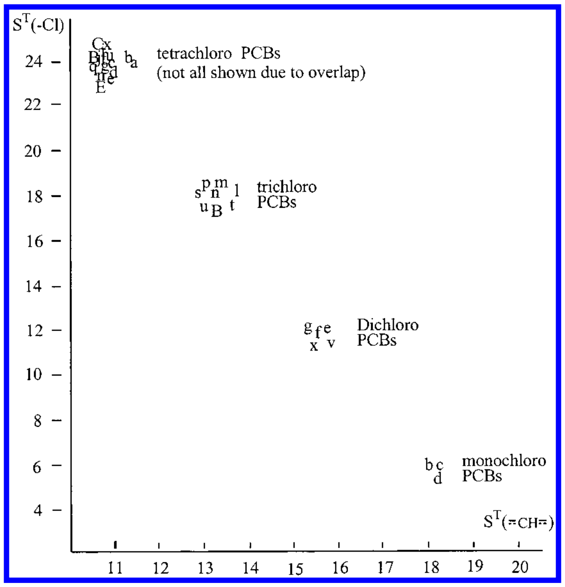


Figure 2. Distribution of monochloro-, dichloro-, trichloro-, and tetrachloro-PCBs in atom-type E-state space, based on two indices: chlorine [$S^T(-Cl)$] and aromatic CH, [S^T(⋯CH⋯)].

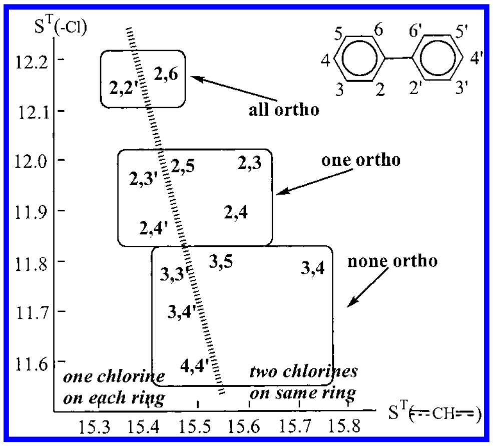


Figure 3. Structurally meaningful distribution of dichloro-PCBs in E-state space as a projection onto two dimensions of E-state space, based on two atom-type indices: Chlorine [$S^T(-Cl)$] and aromatic CH [S^T(⋯CH⋯)]. See text for structure analysis.

a meaningful way in terms of molecular structure and represents chemically relevant information.

The organization of these PCBs in this parameter space is even more evident in the case of trichlorobiphenyls shown in Figure 4. As in Figure 3, the topmost substitution pattern contains chlorines only in the ortho position: 2,6,2′. Further examination of this figure reveals three other groupings indicated by three boxes. As found in the dichloro-PCBs, the bottom box contains PCBs with chlorines most removed from the ortho positions. The intervening boxes contain molecules with chorines closer to the ortho positions—closer in the higher boxes, further in the lower boxes. The number of ortho chlorines progresses smoothly from 3 to 2 to 1 to none from the top box to the bottom box. Within each box the molecules with chlorine atoms closer to the ring juncture in **I** are higher in the box. As can be seen here, the pattern of organization for trichloro compounds is similar to that found for the dichloro compounds. The distinctive character of chlorines found in the ortho position parallels the role of chlorines as described by Seybold for the gas chromatographic retention of PCBs.[17]

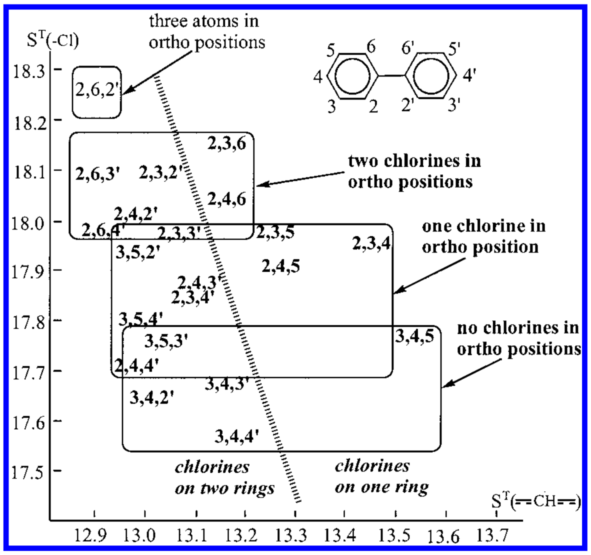


Figure 4. Structurally meaningful distribution of trichloro-PCBs in E-state space as a projection onto two dimensions of E-state space, based on two atom-type indices: Chlorine [$S^T(-Cl)$] and aromatic CH [S^T(⋯CH⋯)]. See text for structure analysis.

A second pattern in the grid of trisubstituted PCBs is found from an examination of the location of compounds in the figure with respect to the number of chlorines on each ring. Those PCBs with all the chlorines on one ring are found to the right of the dashed line in the figure. Those with chlorines on both rings are found to the left. Further, those structures in the left most part have chlorines closest to ortho positions whereas those closest to the line (but still on the left) have chlorines further from the ortho positions. A similar pattern is observed for the right side. This additional complexity is not as common for the simpler dichloro-PCBs because there are fewer possible isomers.

It is clear that the structure pattern is similar in each PCB cluster. The arrangement is based on clear features of chemical structure and readily apparent and easily assessed. This pattern information was not forced into the E-state formalism a priori. It is the consequence of the fundamental definition and information content in the E-state formalism.

Similarity Searching Using the Atom-Type E-States. A commonly held view is that molecules which are similar in structure may have similar properties of comparable magnitude. In the context of biological properties, similarity presages comparable activities as ligands, substrates, inhibitors, or other agents engaged in intermolecular encounters in living systems. In the area of drug design, this similarity is the guiding principle in structure−activity modeling. A mathematical relationship is sought between a measured activity and a molecular property or structure quantitation. The purpose is to predict another structure, representing a candidate molecule, and to shed light on the structure−activity relationship. The rationale is that similarity may portend comparable behavior. The central element in this dialectic is similarity. This concept and associated methodology are explored in this section.

The E-state concept is a prime candidate for the definition of similarity among molecules, molecular fragments, and atoms-in-molecules.[18] One reason for this conjecture is that the E-state analysis produces numerical values which encode the extent of some important attributes. Structure information may be represented at the atom level, the atom-type level, and the bond-type level, providing a broad basis for encoding molecular structure.[1] We show here examples of molecular similarity searches using two drug molecules as references and the atom-type E-state indices for the basis of similarity. The database used is a modification of the Pomona MedChem database, which contains 21 000 molecular structures.[18] The atom-type E-state indices constitute the database along with names and SMILES strings. Euclidean distances were computed between a target drug molecular structure and each molecule in the database. Each atom-type E-state index was converted to a z score: $z_i = (x_i - \mu_i)/\sigma i$; x_i is the ith E-state index, μ_i its mean, and σ_i its standard deviation in the database.[16a] In this manner, each dimension in the structure space is put on a comparable basis of magnitude and spread. The database searching was performed with the FORTRAN program ESEARCHZ developed by Hall. To assess structure similarity, both the Euclidean distance (generalized Minkowski distance)[16b] and the cosine formula[16c] are used.

$$d_{\text{Euclidean}} = [\sum_j (z_{\text{ref}} - z_j)^2]^{1/2}$$

$$\text{cosine} = \sum_j (z_{\text{ref}} z_j)/[\sum_j (z_{\text{ref}})^2]^{1/2}[\sum_j (z_j)^2]^{1/2}$$

The sum is performed over the specified number of dimensions. Structures are ranked on their Euclidean distances. Both Euclidean distance and the cosine are reported in the tables. The structures found closest to the target drug structure are listed in the two tables.

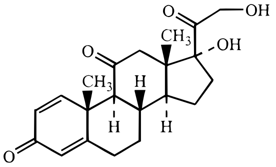
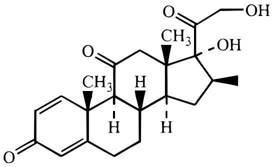
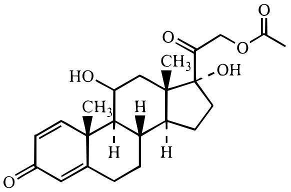
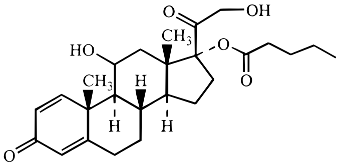
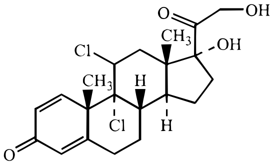
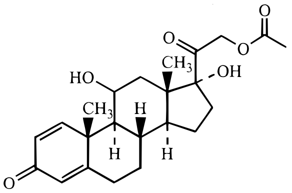
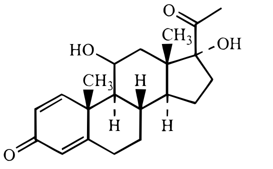
Prednisone. The antiinflammatory agent prednisone is a commercial drug representing this class of drugs where there is a continuing need to find new drugs to deal with inflammation. Using this molecule as a reference, we can search a database described by atom-type indices to identify structures that are similar on the basis of low numerical Euclidean distances. All eight atom-types in prednisone are used in the search: $S^T(-CH_3)$, $S^T(-CH_2-)$, $S^T(>CH-)$, $S^T(>C<)$, $S^T(=CH-)$, $S^T(=C<)$, $S^T(-OH)$, and $S^T(=O)$. Table 3 shows the first seven structures found in the search. Prednisone is found at the distance = 0. All the close structures have the steroid nucleus. The closest structure is very similar, methylprednisone, differing only by a methyl group in the 17-position. It is clear that low Euclidean distance and cosine values near 1.0 indicate similarity of molecular structure. Also note that five of the structures found are drugs, and their trade names are given.

Mefloquine. Mefloquine is a well-known antimalarial drug. Using the same database as in the previous example, compounds similar to Mefloquine are identified using atom-type E-state indices and Euclidean distances for the similarity metric. In the first search for mefloquine-like structures, all 10 atom-types found in mefloquine are used:

$$S^T(-CH_2-), S^T(>CH-), S^T(>C<), S^T(\cdots CH\cdots), S^T(\cdots \overset{|}{C} \cdots), S^T(-NH-), S^T(\cdots N\cdots), S^T(-OH), \text{and } S^T(-F)$$

In Table 4a, the results of the search are revealed. It is clear that the closest compounds have very similar structures.

Table 3. Database Similarity Search with Reference Compound Prednisone (Using All Eight Atom Types in the Structure)[a]

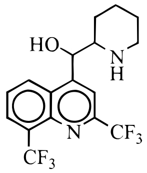
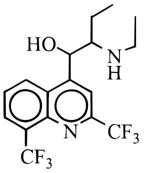
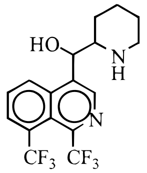
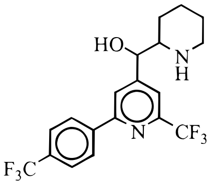
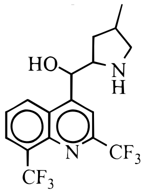
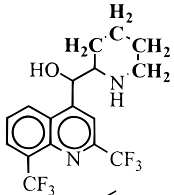
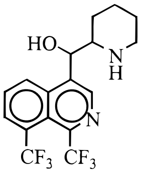
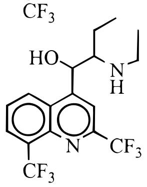
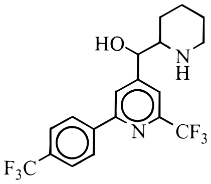
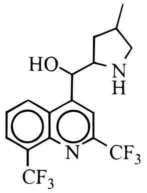
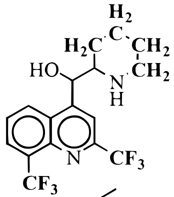
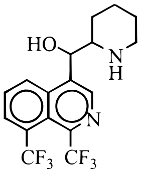
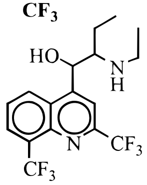
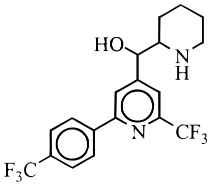
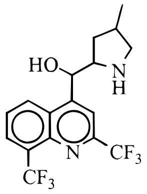
mol struct	Euclidean distance (arbitrary units)	cosine
 Prednisone	reference	1.000
 Meprednisone	0.75	0.989
 Prednisoloneacetate	0.88	0.984
 Prednival	1.18	0.973
 Dichlorisone	1.29	0.964
 6-α-Methylprednisoloneacetate	1.45	0.961
 Deprodone	1.47	0.959

[a] Note: Prednisone found distance = 0.0. Next closest compound at distance = 1.51; 13 more found up to distance = 1.83.

It is known that the methylene groups of the saturated amine ring are not significant for biological activity. A second search was conducted in which only nine atom-type E-state indices were used, leaving out $S^T(-CH_2-)$. The results are shown in Table 4b. It is interesting to note that the first three compounds found are the same but in a different order.

A third search was performed in which only seven atom types were used; the $S^T(-CH_2-)$, $S^T(>C<)$, and $S^T(-F)$ were excluded, based on the idea that these features are not

Table 4. Database Similarity Search with Reference Compound Mefloquine

mol struct		Euclidean distance (arbitrary units)	cosine	mol struct		Euclidean distance (arbitrary units)	cosine
A. Using All 10 Atom-Types in the Structure							
	mefloquine	reference	1.000			0.61	0.998
		0.48	0.999		enpiroline	1.25	0.994
		0.53	0.999				
B. Using 9 Atom-Types in the Structure Excluding $-CH_2-$[a]							
	mefloquine	reference	1.000			0.48	0.999
		0.13	1.000		enpiroline	1.25	0.994
		0.14	1.000				
C. Using 7 Atom-Types in the Structure Excluding $-CH_2-$ and $-CF_3$[a]							
	mefloquine	reference	1.000			0.61	0.998
		0.48	0.999		enpiroline	1.25	0.994
		0.53	0.999				

[a] Note: Mefloquine found distance = 0.0.

essential to biological activity. In this search the same order is obtained for the first three compounds found similar to mefloquine (Table 4c). This result may appear unexpected at first because three atom-types were excluded. However, this finding underscores the reciprocal nature of the E-state formalism. The E-state value for each atom contains information about every other atom in the structure, an indication of the information-field nature of the E-state values. Because of this basic nature of the E-state, the user may formulate approaches to tailor-make a search. Less important or nonessential features may be excluded. In this manner the user has a high degree of control and can be very creative. Structures found in this search method which are apparently less similar to the reference structure may prove useful as new drug candidates.

CONCLUSION

The electrotopological state (E-state) index is shown to contain information reflecting intermolecular accessibility of atoms and groups in a molecule, specifically electron accessibility. This information is encoded into a numerical value reflecting the electronegativity and the topology of each atom. The index for an atom, the electron accessibility at that atom, is sensitive to the electronegativities and topological state of all other atoms in the molecule. This perturbing effect is carried through the network of atoms, described by a chemical graph.

The atom-type E-state indices are defined, for subsets of atoms of the same type, as the sum of the individual E-state, thus making it possible to conduct structure−activity analyses with diverse structures. The E-state atom-type indices may also be used to organize a very diverse database of molecules into a coherent mosaic of molecules with a strong potential for navigating and searching in a logical way. This is demonstrated by the organizing of the PCBs into a database wherein any compound may be found by inspection and the structure of the neighbors inferred. This capability forms the basis of the description of similarity by any criterion that is chosen. This feature is demonstrated by the search for similar compounds in a database, relative to a reference molecule. Thus lead compounds may serve as reference molecules in the search for potentially active congeners or bioisosteres, based upon an exploitable criterion of similarity. These opportunities are made available using the E-state atom-type indices.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, 1999.

(2) Kier, L. B.; Hall, L. H. The Electrotopological State: Structure Modeling for QSAR and Database Analysis. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A., T.; Eds.; Gordon and Breach, Reading, U.K., 1999.

(3) Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(4) Hall, L. H.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−83.

(5) Kier, L. B.; Hall, L. H.; Frazer, J. W. An Index of Electrotopological State for Atoms in Molecules. *J. Math. Chem.* **1992**, *7*, 229−237.

(6) Hall, L. H.; Kier, L. B. The Electrotopological State: An Atomic Index for QSAR. *Quant. Struct-Act. Relat.* **1991**, *10*, 43−48.

(7) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. The E-State Fields. Application to 3D QSAR. *J. Comput. Aided Mol. Des.* **1997**, *10*, 513−520.

(8) Gough, J.; Hall, L. H. Modeling the Toxicity of Amide Herbicides using the Electrotopological State. *Environ. Tox. Chem.* **1999**, *18*, 1069−1075.

(9) Hall, L. H.; Story, C. T. Boiling Point of a Set of Alkanes, Alcohols and Chloroalkanes: QSAR with Atom Type Electrotopological State Indices using Artificial Neural Networks. *SAR QSAR Environ. Res.* **1997**, *6*, 139−161.

(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(11) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−587.

(12) (a) Mulliken, R. S. A New Electroaffinity Scale. *J. Chem. Phys.* **1934**, *2*, 783−793. (b) Hinze, J.; Jaffe, H. H. Electronegativity. I. Orbital Electronegativity of Neutral Atoms. *J. Am. Chem. Soc.* **1962**, *84*, 540−549.

(13) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.

(14) MOLCONN-Z may be obtained from Hall Associates Consulting, 2 Davis St., Quincy, MA; SciVision Inc., 200 Wheeler Street, Burlington, MA 01803; Edusoft, LC, P.O. Box 1811, Ashland, VA 23005; and Tripos, Inc., 1699 South Hanley Rd., St. Louis, MO 63144.

(15) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley-Interscience: New York, 1990.

(16) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press, Wiley: New York, 1987; a, pp 50−51; b, p 52; c, p 54.

(17) Seybold, P. G.; Bertrand, J. A Simple Model for the Chromatographic Retentions of Polyhalogenated Biphenyls. *Anal. Chem.* **1993**, *65*, 1631−1634.

(18) Hall, L. H.; Kier, L. B.; Brown, B. B. Molecular Similarity based on Novel Atom-Type E-State Indices. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1074−1080.

CI990140W