

Molecular Similarity Based on Novel Atom-Type Electrotological State Indices

Lowell H. Hall*

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

Lemont B. Kier and Briscoe B. Brown

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

Received September 11, 1995[⊗]

A molecular similarity measure is developed from a structure representation based on atom level topological and electronic information, specifically the electrotological state indices. For the purpose of creating a structure information space, atom typing is introduced. The classification of each atom in the molecular graph is derived from its valence state. The electrotological state indices for all atoms of the same type in the molecule are summed to create an atom-type E-state index for each atom type. These indices are shown to encode significant electronic and topological information. Further, they are shown to be very useful in the representation of structure information as a basis for molecular similarity judgment and for database characterization, including searching. Examples are given for similarity among a set of simple structures and for database searching using four drug molecules.

INTRODUCTION

There is a long-standing axiom in medicinal chemistry that molecules which are structurally similar, by specified criteria, tend to have related chemical and biological activities. This axiom supports the traditional approaches to molecular modification in synthesis-test-synthesis iterations. It is also the basic principle in drug design based on QSAR modeling. Finally, it is the guiding spirit in chemical database management systems. The principle that similar molecules may elicit similar biological responses is commonly accepted, but there appears to be no widely accepted definition of molecular similarity, either in principle or in practice.^{1–4}

In our approach to this important area of investigation, we appreciate that molecules are complex objects and their interactions are no less complex; hence, similarity is not uniquely defined. Similarity depends upon the set of molecules under consideration as well as the particular purpose for which it is used. A molecule is characterized by its properties; its profile of property values sets it apart from all other molecules. A molecule is also characterized by its structure; it possesses a set of atoms arranged in a unique manner. Yet some molecules may share in common the same set of atoms and possess a narrow range of property values. In that sense those molecules may be said to be similar. T. S. Eliot put it this way: "All cases are unique and also similar". This statement encapsulates both the charm and the challenge of chemistry.

The specific property values for a set of molecules together with their molecular structure elements clearly indicate uniqueness for each molecule. Nonetheless, chemists perceive a similarity among a given set of molecules with respect both to molecular structure and a selected set of property values. Our objective is to develop an appropriate structure representation which reveals the perceived similarity and provides a working basis for application of that similar-

ity. Further, it is desirable to have an approach which intrinsically provides more than one criterion for similarity. Based on our experience over the past two decades in the use of indices derived from chemical topology, we recognize the versatility of the indices in representing molecular structure.⁵ In this paper we pursue the application of one particular set of topological indices, focusing on atom level description, as a basis for molecular similarity.

Willett suggested that there are four types of molecular structure representation: systematic nomenclature, fragmentation codes, line notations, and connection tables.² Chemical topology indices now constitute a fifth representation. It is useful to consider them as an additional category even though such indices are usually obtained, for practical reasons, from a connection table representation directly or as an intermediate. A principal difference between this fifth category and the other four is that the presence/absence type offers little discrimination based on atom environment because of their discrete nature. In contrast, graph theoretic quantities, especially the electrotological state indices, provide characterization based on both the electronic and topological environment of each atom in the molecule. Consequently, there is potential for a similarity measure with greater power of discrimination.

The molecular connectivity χ indices represent the topology of the molecular skeleton and can be used as a basis for similarity along with other indices.⁵ In this paper we develop a structure representation for atom electronic information based on the electrotological state indices which encode the electronic state of each atom (hydride group) in the molecule.⁶ Atom typing (classification) is developed. This atom typing is blended with the electrotological state formalism to provide a descriptive structure space for molecules: atom-type E-state indices. The space is orthogonalized using principal components. Similarity is based on Euclidian distance computed from a reference molecule. Several examples are given to demonstrate the nature of this similarity measure.

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1995.

METHOD

This paper is based on an approach to atom typing coupled with the electrotopological state indices.⁷ Further, database searching using only designated parts of a molecule will be addressed using the atom type E-state indices in a later paper.

Electrotopological State Space. A new paradigm, called the electrotopological state for atom electronic and topological characterization, was introduced by Kier and Hall and was reviewed recently.⁶ For simplicity, these indices are referred to as the E-state. Each atom in the molecular graph is represented by an E-state variable which encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule.

The E-state index for an atom consists of an intrinsic value for that atom (in its valence state) plus a term for its perturbation by all the other atoms in the molecule. The intrinsic state is based on the Kier–Hall electronegativity⁸ and derived from the ratio of that electronegativity to the number of skeletal σ bonds for that atom⁶

$$I = ((2/N)^2 \delta^v + 1) / \delta \quad (1)$$

The symbols δ^v and δ are the molecular connectivity δ values which are given as follows (for first row atoms)

$$\delta = \sigma - h = \text{number of connections in the skeleton} \quad (2)$$

where σ is the number of electrons in σ orbitals; h is the number of hydrogen atoms bonded to the atom.

$$\delta^v = Z^v - h = \sigma + \pi + n - h \quad (3)$$

Z^v is the number of valence electron; π is the number of electrons in π orbitals; n is the number of electrons in lone pairs; N is the principal quantum number of the valence shell for that atom. It can be seen $\delta^v - \delta$ is equal to the number of π and lone pair electrons which Kier and Hall showed is proportional to valence state electronegativity.⁶ The intrinsic state value is large for electronegative atoms and decreases for less electronegative atoms and for atoms with several σ bonds.

The E-state index for atom i , S_i , is defined as follows

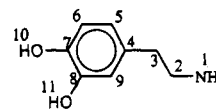
$$S_i = I_i + \Delta I_i \quad (4)$$

$$\Delta I_i = \sum (I_i - I_j) / r_{ij}^2 \quad \text{sum over all atoms} \quad (5)$$

where r_{ij} is the distance between atoms counted as the graph distance (d_{ij}) plus one. As a consequence of this definition, atoms tend to have large positive values for S_i if they possess π and lone pair electrons and are terminal atoms. Atoms which do not have π and lone pair electrons and/or are buried in the interior of the molecule ($\delta \geq 4$) tend to have small or negative E-state values. See Table 1 for a sample calculation and the references cited below.

E-state indices have been used to correlate ¹⁷O NMR frequencies for ethers, aldehydes, and ketones;^{6,9,11} binding of a series of indolealkylamines to 5-HT₂ receptors;⁵ binding of barbiturates to β -cyclodextrin;⁹ inhibition of flu virus by benzimidazoles;¹⁰ receptor binding affinity of β -carbolines;¹² binding of salicylamides to the dopamine D-2 receptor;¹³ and inhibition of MAO by hydrazides.¹² The MAO inhibition study was extended to include careful analysis of the inhibitor

Table 1. An Illustration of the Computed E-State Value for Each Atom and Their Sums for the Atom Types Present in the Molecule and the Atom-Type E-State Indices of the Types Used in This Work



atom no.	atom type symbol	atom electrotopological state index value
1	sNH2	5.304
2	ssCH2	0.546
3	ssCH2	0.716
4	aasC	0.933
5	aaCH	1.738
6	aaCH	1.459
7	aasC	-0.092
8	aasC	-0.087
9	aaCH	1.516
10	sOH	8.931
11	sOH	9.036

atom type E-state symbol	index value
SssCH2	1.262
SaaCH	4.713
SaasC	0.752
SsNH2	5.304
SsOH	17.967

molecules using semiempirical MO computations. The model based on the E-state indices was found to be superior to that based on MO computed charges;¹⁴ the time requirements for the MO study was about 1000 times more than that required for the E-state analysis.

As developed in earlier papers on the E-state, each atom in the molecular graph is represented by an index value. To make the E-state indices more generally applicable to a wide range of molecular structures found in the typical database, we have subsequently developed an atom typing concept.⁷ For our present purposes only a brief discussion will be given.

Atom Classification Scheme and Nomenclature. Each atom in the molecule is classified according to its valence state, including the number of attached hydrogen atoms. Such groupings are sometimes called hydride groups. Such classification is carried out in a modified version of the Molconn-X software.¹⁵ For the atom type indices, the E-state value is computed for each atom; values are summed for all atoms of the same type. Table 2 gives a list of the atom types used in this paper along with the symbols used.

Each atom type E-state symbol is a composite of three parts. The first part is "S" which stands for the sum of E-state values for all atoms of the same type in the molecule. The second part is a string representing the bond types associated with that atom, including "s" for single bond, "d" for double, "t" for triple, and "a" for aromatic. Finally, there is symbol for the set of atoms in the hydride group, such as CH₃, CH₂, OH, Br, or NH. For example, the symbol SssCH₂ stands for the sum of E-state values for all the -CH₂- groups in the molecule; SaaCH stands for the sum of E-state values for the CH groups in an aromatic ring; SsNH stands for the sum of E-state values for secondary amine groups. Table 1 gives the E-state atom values as well as atom type E-state values for an example.

To illustrate how atom type E-state indices represent structure in an information space, consider the alcohols and

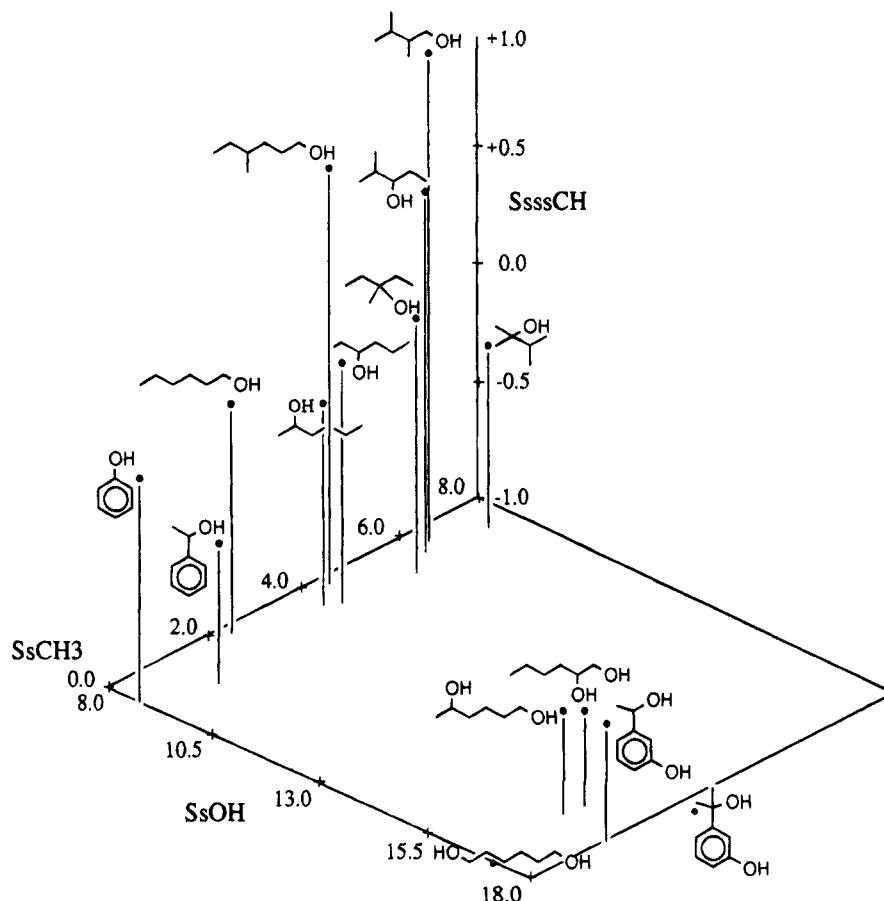


Figure 1. An illustration of the atom-type E-state space for a small set of alcohols and diols using a limited set of indices. The three axes are for the OH group (SsOH), the methyl group (SsCH3), and the methine group (SsssCH). See text for definitions.

diols shown in Figure 1. For this illustration only three atom type E-state indices are used: SsCH3, SsssCH, and SsOH. The SsCH3 variable encodes the electronic and topological state of all the methyl groups in the molecule. Likewise, the SsssCH encodes E-state values for methine groups and SsOH encodes OH groups. The molecules are arrayed in an organized fashion throughout this part of E-state structure space. Not all of the atom type E-state indices present in these molecules are being used in this limited example. Indices for methylene groups as well as aromatic CH and C could also be used but are not included here for the sake of visualization. In this sense, this illustration is a kind of projection in a larger space. Both phenol and 1,6-hexanediol have no methyl groups; hence, their positions correspond to a zero value for SsCH3.

Orthogonalization of E-State Space. For this present investigation we have selected a limited number of atom types for simplicity. Ultimately the list of atom types can be rather large. This issue will be the subject of a subsequent paper. Table 2 gives the list of atom types used in this study. These 26 atom types suffice only to illustrated similarity and database searching. A set of 79 atom types is currently under development along with a large number of associated bond types.

To exploit this structure information more fully, it is useful to make a transformation to an orthogonal space. We used the method of principal components to create an orthogonal space with the minimum number of dimensions and the maximum structure information. To obtain the transformation matrix, we used a database obtained from the Pomona MedChem data set, removing a small number of ionic and

Table 2. List of Symbols for the Atom Types Used in this Paper with the Electropotential State Atom Indices for this Development of Molecular Similarity^{c,d}

symbol	structural formula	symbol	structural formula
SsCH3	-CH ₃	SaaN	..N..
SdCH2	=CH ₂	SsssN	>N-
SsssCH2	-CH ₂ -	SddsN	==N- ^b
SdsCH	=CH-	SsOH	-OH
SaaCH	..CH..	SdO	=O
SsssCH	>CH-	SsO	-O-
SdssC	=C<	SaaO	..O..
SaasC	..C.. ^a	SsF	-F
SsssC	>C<	SsSH	-SH
SsNH2	-NH ₂	SssS	-S-
SsNH	-NH-	SaaS	..S..
SaaNH	..NH..	SsCl	-Cl
SdsN	=N-	SsBr	-Br

^a An aromatic carbon with two aromatic bonds and one single bond.

^b The nitrogen atom in a nitro group. ^c The symbols for bond types are as follows: single, s (-); double, d (=); aromatic, a (..). ^d The initial upper case S stands for sum of the electropotential state indices for all the atoms of the specified type in the molecule.

organometallic compounds to obtain a list of 21 842 structures. To these we added the complete set of acyclic alkanes from ethane through the dodecanes (663 structures), a set of aromatic hydrocarbons including alkylbenzenes, PAHs and alkylated PAHs (515 structures), and a set of highly branched hydrocarbons (290 structures), for a total of 23 310 molecular structures. The file of SMILES strings was run through a modified version of Molconn-X (version 2.0) for the computation of the atom type E-state indices.¹⁵ The structures were processed at a rate of about 1000 per min.

Principal component analysis and computation of Euclidian distances was accomplished using the SAS System.¹⁶ All the computational work was done on a 90 MHz Pentium computer.

An examination of the correlation matrix for the 26 atom type E-state indices indicates that these indices are nearly linearly independent; that is, their pairwise correlation coefficients are very small. More than 80% have $r \leq 0.05$, and only one is greater than 0.30. However, to facilitate computations using the Euclidian distance formula in an orthogonal space, we have converted the set of 26 variables to their principal components.¹⁶ The amount of variance accounted for by each component is rather homogeneous throughout the principal components; the fraction variance ranges from 10% down to 1%. We decided for this paper to use all 26 PCs in the distance computations for similarity estimation and not to reduce the dimensionality by using fewer components. Other similarity measures, such as the Tanimoto coefficient, could be used so that orthogonalization would not be necessary. Such use is under investigation because the meaning of the original atom type E-state indices would remain for purposes of interpretation.

It is also possible, and indeed very powerful, to use a limited set of atom types for a search. For example, one could limit the set of atom types to those associated with a pharmacophore or other molecular fragment of importance. In that case, an approach other than Euclidian distance and orthogonalization will be preferable. This strategy will be discussed in a subsequent paper.

Strategy for Use. For database searching, a simple and chemically meaningful strategy can be used. The structure for a molecule of interest is considered as a reference and its atom-type E-state indices are computed.¹⁵ Then the distance is computed to every molecule in the orthogonalized E-state space using the standard formula for Euclidian distance. Those structures found at small distances are considered to be the most similar in the database. To get some calibration of the computed distance as a similarity measure, the user can compute distances to structures which the user considers similar. Comparison of distances to these compounds of known similarity gives a useful gauge on the meaning of the computed distances for the compounds found in the database.

In the strategy described here, the user does not have to construct queries made from substructures along with relationships among the substructures. A reference molecule or set of molecules, hypothetical or taken directly from the investigation at hand, is all that is needed.

Illustrations of Similarity Use in Database Searching in E-State Space. To illustrate the database searching capabilities of the E-state representation, two types of studies were performed. First several simple molecules were analyzed in order to determine the general features of this similarity analysis. Second, we have selected four drug molecules as examples of somewhat more complex structures. For these drug molecules, each structure is shown in a table along with the molecules found most similar in the database together with the computed similarity distance for each candidate.

The E-state space used here has 26 dimensions, corresponding to the 26 atom type E-state indices used in the space representation (converted to their principal components). The Euclidian distance formula possesses a sum of the squares

of the difference between coordinates for the reference compound, x_{ref} , and each of the molecules in the database, x_i

$$d = [\sum (x_{\text{ref}} - x_i)^2]^{1/2} \text{ sum over 26 PC coordinates} \quad (6)$$

It should be noted here that the similarity distances given in this paper are in arbitrary units. A larger number of atom types can be easily accommodated into this analysis.

The first set of simple molecules used are alkanes, cycloalkanes, and alkyl substituted benzenes. When octane is used as the reference molecule, the closest two molecules are heptane and nonane, both at an (arbitrary) distance of 0.39 and both also unbranched. No branched molecules are found with a distance less than 0.75 because the unbranched alkanes have a different set of atoms types than do the branched isomers. When 3,3,4-trimethylhexane is the reference, seven trisubstituted nonane isomers are found at small distances around 0.04: 2,4,4-trimethylhexane ($d = 0.024$); 2,2,3-trimethylhexane ($d = 0.035$); 2,3-dimethyl-3-ethylpentane ($d = 0.035$); 2,2-dimethyl-3-ethylpentane ($d = 0.039$); 2,2,3-trimethylhexane ($d = 0.040$); 2,2,4-trimethylhexane ($d = 0.040$); and 2,2,5-trimethylhexane ($d = 0.045$). The next closest molecules possess either one more or one fewer carbon atoms and are either trisubstituted octanes or decanes at distances around 0.35. Likewise, when 3,3,4,4-tetramethylhexane is the reference, only tetrasubstituted hexanes or pentanes are found close by, at distances ranging from 0.01 (2,2,3-trimethyl-3-ethylpentane) to 0.06 (2,2,5,5-tetramethylhexane).

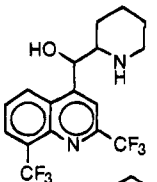
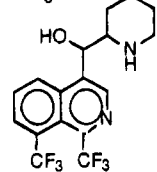
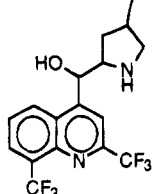
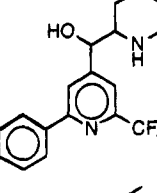
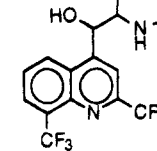
A similar result is found with monocyclic molecules. When cyclohexane is the reference, cycloheptane and cyclopentane are the nearest neighbors at 0.40. When 2-methylethylbenzene is the reference, the *meta* and *para* isomers are the closest neighbors at 0.05 and 0.07, respectively; molecules with one more carbon atom, such as 1,2-diethylbenzene, are around 0.31. For 2-pentanol as reference, 3-pentanol is at 0.08, 3-hexanol is at 0.26, and 2-butanol at 0.31. The nearest neighbors to 1,2-dichloroethane are dichloromethane at 0.27 and 1,3-dichloropropane at 0.32 and 1,2-dichloropropane at 0.59. When 4-ethylphenol is the reference, the 3- and 2-isomers lie at 0.05 and 0.13, respectively. Methyl and propyl substituted compounds lie at distances around 0.3. Similarly for *m*-chlorophenol; the *para* isomer lies at a distance of 0.08 and *ortho* at 0.15; *m*-chlorobenzylalcohol is at a distance of 0.49.

The results for the four drug molecules are shown in Tables 3–6. The first example for searching a database is the drug molecule mefloquine. The first candidate molecule shown in Table 3 is rather similar to the reference, mefloquine; three other candidates are also given. The second example is meperidine for which several rather similar candidates were found and are shown in Table 4. In Table 5 five candidates are listed when niridazole is the reference molecule; the first candidate is somewhat similar. Finally, the fourth reference drug molecule is minoxidil, Table 6. One molecule is found to be somewhat similar; three other candidates are also listed.

DISCUSSION

The electrotopological state encodes both electronic and topological information for each skeletal atom (hydride

Table 3. Molecules Found at Various Euclidian Distances from the Reference Compound, Mefloquine, Based on Their Set of Atom-Type E-State Indices

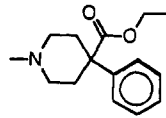
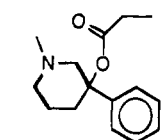
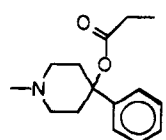
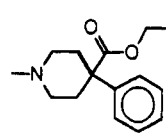
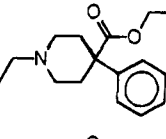
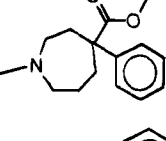
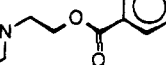
molecular structures	distance (arbitrary units)
	0.23
	
	0.74
	0.75
	1.13

^a This list is for purposes of illustration and was arbitrarily terminated at a distance of 1.13. ^b The distance is in arbitrary units, computed as the Euclidian distance based on the 26 principal components in atom-type E-state space. See text.

group) within the molecule. Highly electronegative atoms have large intrinsic state values and tend to have positive perturbations, according to eq 4. As a result, their E-state values tend to increase above their intrinsic state values. Atoms which are on the periphery or mantle of the molecule also tend to increase in E-state value. Conversely, atoms which are bonded to three or more other atoms tend to have smaller or negative E-state values. For this reason the E-state can be thought of as a measure of electron accessibility.⁶

Because of the nature of the E-state, there is a clear categorization of atoms in their various valence states with a range of E-state values within each atom type. In Figure 1 this arrangement in E-state space is illustrated. Primary monohydric alcohols tend to have SsOH values around 8.3 to 8.5; secondary alcohol values range from 8.5 to 8.9; and tertiary lie above 9.0. The values for diols are, of course, about doubled but there is also a range depending upon bonding environment: 16.60 to 18.44 for SsOH in this example. In a similar manner, the other atom type E-state values cover a range of values which is dependent upon the topological environment of the atoms. For methyl groups, SsCH₃ is greater for methyl bonded to a methine and greatest when on a quaternary carbon; it is of intermediary value

Table 4. Molecules Found at Various Euclidian Distances from the Reference Compound, Meperidine, Based on Their Set of Atom-Type E-State Indices

molecular structures	distance (arbitrary units)
	0.12
	
	0.14
	0.26
	0.31
	0.31
	0.53

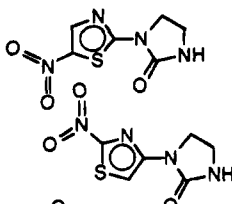
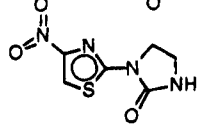
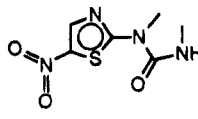
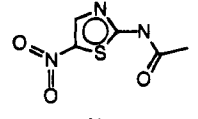
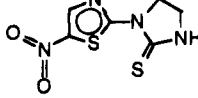
^a This list is for purposes of illustration and was arbitrarily terminated at a distance of 0.53. ^b The distance is in arbitrary units, computed as the Euclidian distance based on the 26 principal components in atom-type E-state space. See text.

when the methyl is on an aromatic carbon. This is an important advance over the use of substructure fragments or keys which operate only on a presence/absence basis. Not only the presence or quantity for the atom type is encoded but also the electronic quality for that atom in that particular molecule.

The nature of the atom type E-state indices indicates that molecules will be distributed in a meaningful manner in the E-state space. The stratified set of values for E-state indices for various atom types reinforces the atom typing to create an arrangement of molecules with meaningful chemical information. Terminal atoms such as -F, =O, and -OH have large values; groups such as -O-, -NH- have intermediate values; >CH-, >C<, and =C< have small and negative values. Location in the E-state space, thus, has a chemical interpretation.

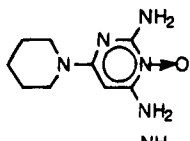
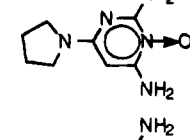
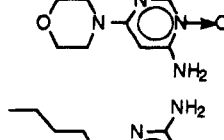
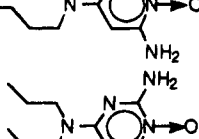
From the analysis on the several simple reference molecules, some general features of the atom-type E-state similarity analysis can be discerned. First, nearest neighbors tend to have the same or very similar skeletal structure. For alkenes, cyclic or acyclic, the nearest neighbors to an unbranched reference molecule are also unbranched. This effect is an illustration of the powerful role of atom typing.

Table 5. Molecules Found at Various Euclidian Distances from the Reference Compound, Niridazole, Based on Their Set of Atom-Type E-State Indices

molecular structures	distance (arbitrary units)
	0.25
	0.85
	0.99
	1.20
	1.50

^a This list is for purposes of illustration and was arbitrarily terminated at a distance of 1.50. ^b The distance is in arbitrary units, computed as the Euclidian distance based on the 26 principal components in atom-type E-state space. See text.

Table 6. Molecules Found at Various Euclidian Distances from the Reference Compound, Minoxidil, Based on Their Set of Atom-Type E-State Indices

molecular structures	distance (arbitrary units)
	0.33
	1.23
	1.31
	1.33

^a This list is for purposes of illustration and was arbitrarily terminated at a distance of 1.33. ^b The distance is in arbitrary units, computed as the Euclidian distance based on the 26 principal components in atom-type E-state space. See text.

Branched isomers have a different set of atom type from those in the unbranched alkane. Branched alkanes are the

nearest neighbors to a branched alkane; further, the nearest neighbors have the same branching types. For example, when the reference has two four-way branches (e.g., 3,3,4,4-tetramethylhexane), the nearest neighbor also has two four-way branch points. The same is also true for cyclic compounds as in the cyclohexane and benzene examples. When heteroatoms are involved, the nearest neighbors have the same set of heteroatoms as shown in the alcohol, phenol, and chlorine-containing examples.

Some further observations can be made. Nearest neighbors tend to have the same number of atoms. Molecules which are perceived to be very similar lie at small distances, generally less than 0.10. Molecules which are similar but which possess one more (or one less) carbon atom tend to lie at somewhat greater distances, usually exceeding 0.25 to 0.30. Isoskeletal molecules with differences in heteroatoms lie at even greater distances.

The four examples of database searching further support the significance of the structure information encoded in the atom-type E-state space. In the first example, mefloquine in Table 3, the first candidate molecule found in the database is somewhat similar to the reference. The difference in structure lies in the reversal of the positions of the ring nitrogen atom and the CF₃ substituent. Changing the six-membered ring to a five-membered ring decreases the similarity considerably for the second candidate. The third candidate differs in that the two fused six-membered rings are replaced by a biphenyl type skeleton. The second drug example, meperidine, shows several candidates which are rather similar in Table 4. The first two show the ester group in reversed orientation and the third has a propyl instead of an ethyl group. The last one shown possesses a diethylamino group instead of the six-membered ring amine and a different arrangement of the ester group with respect to the amine.

For the third searching example, niridazole (Table 5), the difference between the first candidate and the reference lies in the position of the sulfur and the nitro group. The distance 0.25 indicates similarity but not a high degree of similarity. The third and fourth candidates are rather less similar because the five-membered ureido ring is open.

For the last example, minoxidil (Table 6), all the molecules found have one portion of the molecule in common, the diaminopyrimidine nucleus. The molecule in the database found most similar to the reference minoxidil differs in that the six-membered ring is replaced by a five-membered ring. The distance 0.33 indicates that the molecule is somewhat similar. The remaining candidates all differ in the substituted amine portion. The second most similar candidate possesses a morpholino ring and is judged by the distance, 1.23, not to be highly similar.

CONCLUSIONS

The electrotopological state indices have demonstrated considerable usefulness in the establishment of QSAR equations. The ability to focus on individual atoms has provided significant utility in their applicability. The fact that they encode important electronic and topological information invests in them the ability to portray significant pharmacological information for database characterization.

The addition of atom typing extends the usefulness of the E-state indices. This present application deals with molecular similarity. The examples given demonstrate that the atom

type E-state representation provides an arrangement of molecules in a space which gives organization to the electronic character of the molecules. Molecules with similar sets of atoms are close together in the E-state space, yet there is gradation among the molecules according to the variation in the computed E-state indices. Thus, close proximity in E-state space means similarity of structure; separation provides for discrimination based upon differing electronic and topological character.

REFERENCES AND NOTES

- (1) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley Interscience: New York, 1990.
- (2) *Similarity and Clustering in Chemical Information Systems*; Willett, P. John Wiley: Letchworth, England, 1987.
- (3) Rouvray, D. H. Definition and Role of Similarity Concepts in the Chemical and Physical Sciences. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 580–586.
- (4) Heller, S. R. Similarity in Organic Chemistry: A Summary of the Beilstein Institute Conference. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 578–589.
- (5) Basak, S.; Bertelsen, S.; Grunwald, G. D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 270–276.
- (6) Kier, L. B.; Hall, L. H. An Atom-Centered Index for Drug QSAR Models. In *Advances in Drug Design*; Testa, B., Ed.; Academic Press: 1992; Vol. 22.
- (7) Hall, L. H.; Kier, L. B. Electrottopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* in press.
- (8) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, 70, 583–589.
- (9) Kier, L. B.; Hall, L. H. An Electrottopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, 7, 801.
- (10) Hall, L. H.; Mohnney, B.; Kier, L. B. The Electrottopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 76.
- (11) Hall, L. H.; Kier, L. B. An Index of Electrottopological State for Atoms in Molecules. *J. Math. Chem.* **1991**, 7, 229.
- (12) Hall, L. H.; Mohnney, B.; Kier, L. B. The Electrottopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat.* **1991**, 10, 43.
- (13) Hall, L. H.; Kier, L. B. Binding of Salicylamides: QSAR Analysis with Electrottopological State Indexes. *Med. Res. Rev.* **1992**, 2, 497–502.
- (14) Hall, L. H.; Mohnney, B. K.; Kier, L. B. Comparison of Electrottopological State Indexes with Molecular Orbital Parameters: Inhibition of MAO by Hydrazides. *Quant. Struct.-Act. Relat.* **1993**, 12, 44–48.
- (15) The program Molconn-X was used for computation of electrottopological state indices. Contact author L. H. Hall for information.
- (16) SAS Institute, Cary, NC.

CI950093H