"Définition et exploitation d'une banque de données orientée vers Les propriétés radioprotectrices de composés chimiques". *Automatisme* **1975**, 320-328.
(26) Dubois, J. E.; Bonnet, J. C.; Goldwaser, D.; Attias, R. "The DARC System: A Chemical Information System Based on the Topological Encoding of Chemical Compounds". *Proceedings of Eurim II*; Batten, W. E., Ed.; 1976; pp 135-144.
(27) Dubois, J. E.; Bonnet, J. C. "The DARC Pluridata System: ^{13}C-NMR DATA Bank". *Anal. Chim. Acta* **1979**, *112*, 245-252.
(28) Chretien, J. R.; Szymoniak, J.; Dubois, J. E.; Poirier, M. F.; Garreau, M.; Deniker, P. *Eur. J. Med. Chem.-Chim. Ther.* **1985**, *20*, 315-325.
(29) Dubois, J. E.; Carabedian, M.; Dagane, I. "Computer Aided Elucidation of Structures by Carbon-13 NMR. The DARC-EPIOS Method: Characterization of Ordered Substructures by correlating the Chemical Shifts of Their Bonded Carbon Atoms". *Anal. Chim. Acta* **1984**, *158*, 217-233.
(30) EURECAS is the commercial name of the CAS file handled by DARC structure management system since 1979 (CNIC and Telesystem marketing).
(31) Gordon, J. E. "Chemical Inference. 2. Formalization of the Language of Organic Chemistry: Generic Systematic Nomenclature". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 81-92.

Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors

RAMASWAMY NILAKANTAN,* NORMAN BAUMAN,* J. SCOTT DIXON,† and R. VENKATARAGHAVAN

Lederle Laboratories, Pearl River, New York 10965

Received July 22, 1986

A new molecular descriptor, the *topological torsion* (TT), is described for use in statistical SAR studies. The TT consists of four consecutively bonded non-hydrogen atoms along with the number of non-hydrogen branches. This descriptor is essentially the topological analogue of the basic conformational element, the torsion angle. A comparative study of this descriptor and the *atom-pair* descriptor (Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64-73) using the trend vector and similarity probe methods is presented. These methods are described in detail in Carhart et al. The atom-pair and TT descriptors capture and magnify distinct aspects of molecular topology; judicious use of both of these descriptors could significantly enhance the hit rate in routine screening programs in the pharmaceutical industry.

INTRODUCTION

Several different molecular descriptors have been used in structure–activity (SAR) studies. The interest of the medicinal chemist is in capturing molecular features responsible for pharmacological activity. Although a specific three-dimensional arrangement of atoms may be necessary for activity, the features essential for activity are often found in the topological description of the molecule. Indeed, the possible three-dimensional conformations and pharmacological activity are implicit in the topological description if only we knew how to extract them.

A search of the literature shows essentially two types of molecular descriptors. The first is "holistic" in the sense that the descriptor is *one number*, usually representing some important physical property of the molecule as a whole. Some examples of such descriptors are the estimated 1-octanol–water partition coefficient[1] and the shape index of Kier.[2] The molecular shape indices of Hopfinger[3-5] can also be included in this category. These descriptors are either measured or calculated algorithmically and have been used with considerable success in SAR studies. A drawback, however, is that distinct pieces of information about the molecule, such as the constituent atom types, the bond types, and the pairwise connections, are not preserved. This tends to restrict their application to series of molecules that have a high degree of structural similarity.

The second category of descriptors consists of several distinct pieces of information strung together and may include such things as atomic species, bond types, and connectivity of pairs of atoms. Examples in the literature include the AA (augmented atom) and gAA (ganglia-augmented atom) descriptors of Hodes and co-workers,[6-8] the interactively selectable descriptor set used by Varkony and co-workers[9] for examining structural similarities among compounds, and the triplet ganglia fragments (a unit of three connected non-hydrogen atoms together with the terminal bonds) used in structure–activity studies by Tinker.[6,10] Another example of descriptors of this category are the linear subfragment descriptors of Klopman,[11] which are all the linear subfragments of a molecule that contain from 3 to 12 heavy atoms.

In our laboratory, a simple descriptor of the latter type, called the *atom pair*, has been used successfully in structure–activity studies.[12] The atom pair is defined as a substructure composed of two non-hydrogen atoms and an interatomic separation measured in bonds along the shortest path connecting the two atoms. The description includes the number of heavy-atom connections and the number of π electron pairs on each atom. The atom-pair descriptor corresponds to a clearly identifiable structural feature and is sufficiently easy to calculate to allow handling of large numbers of structures. The atom-pair descriptor can capture possible long-range correlations between atoms in active molecules. In this paper, we propose a new short-range descriptor that is intended not to replace the atom pair but to complement its predictive power. Like the atom pair, it will be found to correspond to clearly identifiable molecular features and to be easy to calculate. We term this the *topological torsion* (*TT*) descriptor.

DEFINITION

We define a topological torsion as a linear sequence of four consecutively bonded non-hydrogen atoms, each described by its atomic type, the number of non-hydrogen branches attached to it, and its number of π electron pairs. For compactness in coding, the number of branches excludes those that go to make the torsion itself. Thus, for the two end atoms of the torsion

* Address correspondence to these authors.
† Present address: Smith Kline and French Laboratories, L-940, Philadelphia, PA 19101.

0095-2338/87/1627-0082$01.50/0

the number of branches is calculated as the total number of branches minus 1, and for the two central atoms this number is calculated as the total number of branches minus 2. Schematically the TT can be illustrated as

(NPI-TYPE-NBR)–(NPI-TYPE-NBR)–
(NPI-TYPE-NBR)–(NPI-TYPE-NBR)

where NPI indicates the number of π electrons on each atom, TYPE indicates the atomic species, and NBR is the number of non-hydrogen branches calculated as described above.

It will be noticed that the TT descriptor is related to, and is an extension of, the triplet ganglia fragments of Hodes and co-workers.[6,10] The linear subfragment descriptors of Klopman[11] also include 4-atom subfragments which are related to the TT descriptor.

SOME HEURISTIC CONSIDERATIONS

The topological torsion descriptor was inspired by the basic conformational element, the torsion angle. In conformational analysis, a molecule is described in terms of its bond lengths (defined by two adjacent bonded atoms), its bond angles (defined by three consecutively bonded atoms, and its torsion angles (defined by four consecutively bonded atoms). Thus, the rationale for defining the TT descriptor as we have done is that the torsion angle (defined by four consecutively bonded atoms) is the minimal structural unit in terms of which the conformation of a molecule can be completely described. The 3-dimensional structure of a molecule can be completely built by using a series of torsions as basic building blocks. Here we use a *topological analogue of the torsion.*

The use of the number of π electrons in the descriptor in place of explicit bond types simplifies the treatment of resonance forms by making the bond type a property of the atoms to which it is attached. In benzene, for example, directly encoding the bond types in the TT descriptor would result in two different TT types in the molecule. On the other hand, coding the π electrons on each atom of the TT implicitly encodes the bonds, making all the descriptors in benzene equivalent.

It should be noted that unlike the atom pair, the TT is not a long-range descriptor. A small change in one part of a large molecule does not affect the TT's in a distant part of the molecule. In contrast, a change of a single atom in a molecule alters all atom pairs involving that atom. In real situations we deal with small biomolecules and seldom encounter very long range correlations. Thus the TT descriptor is expected to be a useful adjunct to the atom-pair in statistical SAR studies.

METHOD AND RESULTS

We carried out a comparative study of the TT descriptor, the atom-pair descriptor, and the augmented-atom (AA) descriptor of Hodes using the similarity probe and trend vector analysis, which have been described in detail in an earlier paper from our laboratory.[12] Although these techniques were developed with the atom-pair descriptor in mind, there is nothing in them that is specific to that descriptor, and they are of broad applicability as indicated in ref 12. A brief description of them follows.

The similarity probe is used to select molecules topologically similar to a given molecule in that they have a large proportion of their descriptors in common with it. Pharmacological activity is not considered. The probe molecule is resolved into its constituent descriptors; then from a large database the set of compounds which have the highest proportion of descriptors in common with the probe molecule are selected. The similarity score is calculated by using the formula

$$S = 2D_{ij}/(d_i + d_j)$$

where d_i and d_j are the numbers of distinct descriptors in structures i and j, respectively, and D_{ij} is the number of descriptors the two structures i and j have in common. Trend vector analysis uses a set of compounds together with their pharmacological activities. The compounds are resolved into their constituent descriptors and represented as points in a high-dimensional space, where each dimension represents one kind of descriptor. The presence or absence of a particular type of descriptor in a molecule is indicated by a coordinate of 0 or 1 along the relevant axis, and the points representing compounds may be thought of as labeled by their activities.

If any real structure–activity correlation exists among this set of compounds, there may be a clustering of those of low activity and those of high activity. A vector, called the trend vector, is drawn pointing from the inactive cluster toward the active cluster. Specifically, it is calculated analogously to the dipole moment from the formula

$$T = (1/N)\sum_i (a_i - A)S_i$$

where N is the number of structures, a_i is the activity of structure i, A is the average activity, and S_i is a vector that represents the list of descriptors in structure i.

The statistical significance of the trend vector is determined by a randomization technique, in which the pharmacological activities of the compounds are randomly reassigned to the "wrong" structures and a spurious trend vector is constructed. This is repeated, say, 40 times. If the real trend vector is significantly longer than the mean of the spurious trend vectors (measured in units of the standard deviation of the length of the spurious vectors), then the trend vector is considered to express meaningful information.

Perception and Coding of the TT. The algorithm used to perceive and code the TT's in a molecule is straightforward and was readily coded in FORTRAN, BCPL, and C. Starting from a connectivity table, the algorithm finds all the possible TT's by looping first over all the atoms and then over three successive levels of branching. Checks are made to assure that the atoms in the TT quartet are distinct and that the same TT is not counted twice in opposite directions.

The types of the four atoms are recorded. Only 13 common atom types are distinguished, all others being lumped together as a fictitious element Y. The number of π electrons (NPI) on each atom of the quartet is calculated from the bond types. The three pieces of information concerning each of the four atoms (viz., the atom type, the number of π electrons, and the number of non-hydrogen-atom branchings) are packed into a 32-bit word. Each atom-type field occupies 4 bits; the number of π electrons and the number of non-hydrogen-atom branchings each occupy 2 bits. To prevent the same TT *type* from being coded in two different ways, a canonical packing scheme is used.

Perception and Coding of the AA Fragments. We wrote a similar subroutine to encode the Hodes AA fragments. Although the original Hodes AA descriptor uses explicit bond types, we again used the number of bonding π electrons on each atom, for the same reason as explained above.

Similarity Probe. The similarity probe method was applied by using atom-pair as well as TT descriptors to select the first 200 compounds "most similar" to nicotine from the FCD (Fine Chemicals Directory) database.[13]

The distribution of scores of the first 1000 compounds selected by the two methods are shown in Figures 1 and 2. As can be seen from the charts, the shapes of the two distributions are very similar. However, there are differences in the actual scores. The similarity score drops off more steeply in the case of the TT descriptor. Hence the similarity scores obtained by using the two different descriptors cannot be compared directly. A comparison of the lists of similarity scores obtained by the

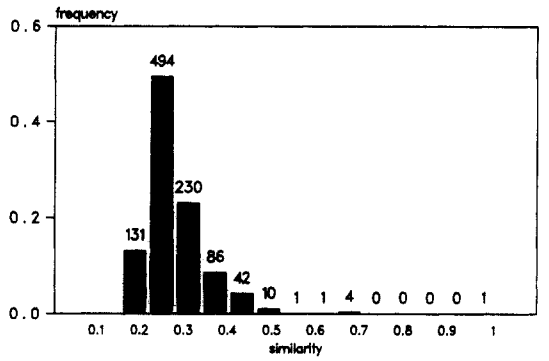

Figure 1. Frequency histogram of atom-pair similarities. The first 1000 compounds most similar to nicotine (on the basis of the atom-pair descriptor) were selected from the FCD database. The histogram shows the distribution of similarity values among these 1000 compounds.

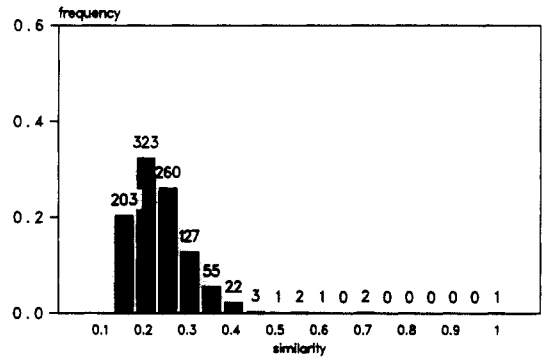

Figure 2. Frequency histograms of torsion similarities. The first 1000 compounds most similar to nicotine (on the basis of the torsion descriptor) were selected from the FCD database. The histogram shows the distribution of similarity values among these 1000 compounds.

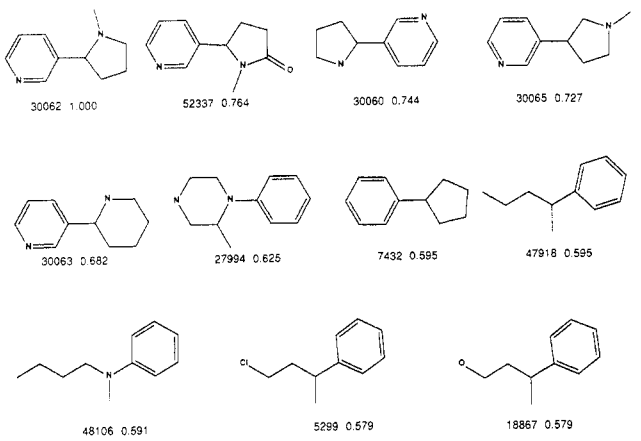

Figure 3. Results of similarity probe on the FCD database using atom-pair descriptors; the probe molecule was nicotine. The first 11 compounds having the highest similarity to nicotine are shown.

two methods shows that a similarity score of 0.65 obtained by the atom-pair method is, in this case, roughly equivalent to a score of 0.50 obtained by the TT method.

The first 11 compounds most similar to nicotine obtained by using atom pairs and TT's are shown, together with their similarity scores, in Figures 3 and 4. The first compound is nicotine, and it naturally has the maximum possible score of 1. Further, it can be seen that the first few compounds are the same in both sets. Beyond this, however, there are differences. For example, FCD 30298 was selected by the TT method and not by atom pairs. It can be seen that the cyclic analogue of this molecule is closely related to nicotine. Another case is FCD 52339, which is a derivative of nicotine. The atom-pair method did not select it because the additional atoms

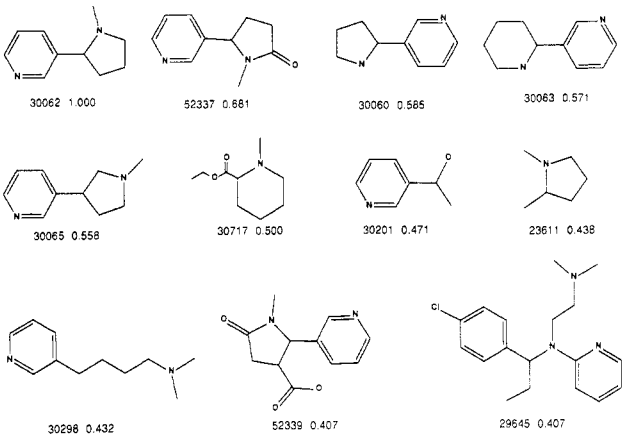

Figure 4. Results of similarity probe on the FCD database using torsion descriptors; the probe molecule was nicotine. The first 11 compounds having the highest similarity to nicotine are shown.

Table I. Correlation Coefficients of Observed vs. Predicted Activity Using Atom Pairs, TT's, and Hodes AA Fragments

training set	test set	descriptor	corr coeff
A	A	atom pairs	0.272
		AA fragments	0.257
		TT	0.338
A	B	atom pairs	0.232
		AA fragments	0.151
		TT	0.278
B	B	atom pairs	0.272
		AA fragments	0.141
		TT	0.443
B	A	atom pairs	−0.008
		AA fragments	−0.008
		TT	0.210

on the 5-membered ring contribute a large number of atom pairs not found in nicotine.

Trend Vector Analysis. Trend vector analysis was applied by using atom pairs, TT's, and AA fragments to a dataset of 4564 compounds tested in the laboratory for binding to the nicotinic receptor.[14] This set was randomly divided into two approximately equal sets, A and B. Trend vectors were constructed as explained above, using the sets A and B separately as training sets. Each of these trend vectors was used to predict the relative activities of both the training set and the test set. For determining prediction rates, the observed activities were converted to all-or-none (0 or 1) values by employing a cutoff value given by the experimenters. The predicted activities were also converted to 0's and 1's by applying a cutoff that resulted in a number of actives equal to that in the observed data. The correlation coefficient between the observed and predicted activities was computed in all cases. The results are shown in Table I.

Not surprisingly, each trend vector was a better predictor of its training set than of the test set. The only exception to this is the case of the AA descriptor where the correlation coefficient for prediction on the training set is 0.141 while that for prediction on the test set is 0.151. This difference, however, is small, and we consider it to be due to random fluctuations. It can be seen that both atom pairs and TT's do significantly better than the AA descriptor. Further, the TT descriptor seems to be marginally better than the atom-pair descriptor.

In the pharmaceutical industry one might be interested in questions of the type, "if we screened only 5% of the compounds in the database in accordance with a theoretical ranking given by a trend vector, what fraction of the actives would be found?". Table II gives a posteriori answers to such questions for atom pairs, TT's, and AA fragments when 5%,

Table II. Percent Actives Discovered Out of a Total of 43 Actives in the Database

level of screening, %	descriptor used	correct predictions on training set, %	correct predictions on test set, %
5	atom pairs	46.5	32.6
	AA fragments	20.9	0.0
	TT	62.8	41.9
10	atom pairs	60.5	46.5
	AA fragments	32.6	27.9
	TT	69.8	46.5
20	atom pairs	67.4	58.1
	AA fragments	44.2	34.9
	TT	72.1	58.1

10%, and 20% of all the compounds in the database are screened.

Again it is clear that the atom pairs and TT's perform significantly better than the AA descriptor. It can also be seen from Table II that at low levels of screening the TT descriptor appears to do better than atom pairs. As we increase the number of compounds screened, all methods of selection converge, and in the extreme case when we screen all the compounds (100% level of screening), we will naturally discover all the actives. Thus, in order to find the relative efficacies of two methods of selection, one should compare the methods at a low level of screening. In Table II we see that at the 5% level of screening TT's perform better than atom pairs while at higher levels of screening both the descriptors do equally well, particularly on the test set.

DISCUSSION AND CONCLUSIONS

The topological torsion, a new descriptor for use in structure–activity studies, has been described, and an algorithm given for its calculation. The TT descriptor is closely related to the triplet ganglia of Hodes and co-workers.[6,10] The TT is slightly more descriptive and includes the number of π electrons on each atom and the branchings on each of the four atoms constituting the TT. We have compared it with both the atom pair and the AA descriptors.

Similarity probe calculations done by using atom pair and TT descriptors give slightly different results. Atom pairs are generally sensitive to small changes even in large molecules, as changing only one of n atoms changes all the $n - 1$ atom pairs involving that atom and deleting one atom changes the distances in all the atom pairs that span it. In other words, the effect of changing a single atom in a molecule is dependent on the total number of atoms in the molecule. TT's, on the other hand, are local, and the effect of changing a single atom in a molecule is independent of the total number of atoms in the molecule. The actual number of TT's altered as a result of altering a single atom depends on the local degree of branching of the molecule.

Thus the two descriptors capture and magnify different aspects of molecular topology. Neither one is necessarily superior to the other. In order to fully exploit the predictive capabilities of these two descriptors, one might merge the lists of compounds obtained by using the two descriptors, reducing the possibility of missing potentially interesting compounds.

Trend vector analysis on a set of 4564 compounds from the nicotinic binding database indicates that the TT descriptor performs slightly better than the atom-pair descriptor and considerably better than the Hodes AA descriptor. Trend vectors constructed by using either atom-pair or TT descriptors were predictive in that they gave results positively correlated with activity when tested on compounds other than those used in their construction. Although the correlation coefficients are numerically small, they are quite significant. It must be recalled that random screening usually has a hit rate of less than 1%; a system of prediction that is wrong 95% of the time represents an enormous improvement over random screening as it has the potential of compressing several years' work into one.

The preceding results clearly demonstrate that atom pairs as well as TT's are useful topological molecular descriptors in statistical SAR studies aimed at improving the hit rate in routine screening programs in the pharmaceutical industry. It cannot be claimed at this point that the TT descriptor is superior. The purpose of this descriptor is simply to complement the type of information captured by the atom-pair descriptor for structure–activity studies. Indeed, it may well turn out that in another example the atom pair is a more appropriate descriptor and gives better results.

The TT descriptor can be refined by introducing a "rigidity" weighting as follows. If the central bond of a torsion is a double or triple bond or an aromatic bond (and hence nonrotatable), we could give it a higher weight than a torsion with a single (and hence rotatable) central bond. The thought behind such a weighting scheme is the presumption that biological activity arises from some specific three-dimensional shape of the molecule. Since nonrotatable bonds confer rigidity on a molecule and fix it into a particular shape, it would be reasonable to weight these rigidifying features high. We plan to try out each enhancement of the TT descriptor in our future studies.

ACKNOWLEDGMENT

This work was supported in part by the U.S. Army Medical Research and Development Command, Contract No. DAMD17-84-C-4111.

REFERENCES AND NOTES

(1) Martin, Y. C. "Classical Concepts of Relationships between Physical Properties and Biological Potency". In *Quantitative Drug Design*; Gruenwald, G. L., Ed. Marcel Dekker: New York, 1978; Vol. 8, Chapter 1.
(2) Kier, L. "A Shape Index from Molecular Graphs". *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109–116.
(3) Hopfinger, A. J. "A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based upon Molecular Shape Analysis". *J. Am. Chem. Soc.* **1980**, *102*, 7196–7206.
(4) Hopfinger, A. J. "Theory and Application of Molecular Potential Energy Fields in Molecular Shape Analysis: A Quantitative Structure–Activity Relationship Study of 2,4-Diamino-5-benzylpyrimidines as Dihydrofolate Reductase Inhibitors". *J. Med. Chem.* **1983**, *26*, 990–996.
(5) Walters, D. E.; Hopfinger, A. J. "Case Studies of the Application of Molecular Shape Analysis to Elucidate Drug Action". *THEOCHEM* **1986**, *134*, 317–323.
(6) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. "A Statistical-Heuristic Method for Automated Selection of Drugs for Screening". *J. Med. Chem.* **1977**, *20*, 469–475.
(7) Hodes, L. "Selection of Molecular Fragment Features for Structure–Activity Studies in Antitumor Screening". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132–136.
(8) Hodes, L. "Computer-Aided Selection of Compounds for Antitumor Screening: Validation of a Statistical-Heuristic Method". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128–132.
(9) Varkony, T. H.; Shiloach, Y.; Smith, D. H. "Computer-Assisted Examination of Chemical Compounds for Structural Similarities". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 104–111.
(10) Tinker, J. "Relating Mutagenicity to Chemical Structure". *J. Chem. Inf. Comput. Sci.* **1981**, *21* 3–7.
(11) Klopman, G. "Artificial Intelligence Approach to Structure–Activity Studies. Computer-Automated Structure Evaluation of Biological Activity of Organic Molecules". *J. Am. Chem. Soc.* **1984**, *106*, 7315–7321.
(12) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. "Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(13) FCD (Fine Chemicals Directory) Copyright 1983, 1984, 1985, by Fraser Williams (Scientific Systems) Limited and Molecular Design Limited.
(14) Kukel, C.; Jennings, K., personal communication.