Pharmaceutical Research, Vol. 6, No. 6, 1989

Report

Differential Molecular Connectivity in Data-Base Fragment Searching

Lemont B. Kier[1,3] **and Lowell H. Hall**[2]

Received July 26, 1988; accepted January 30, 1989

A general scheme is described in which molecular fragments are coded from molecular connectivity values. Specifically a fragment is described by the difference between a simple connectivity index of a certain order and the valence connectivity index of the same order. This numerical value is then used to search for that particular fragment among stored fragment values associated with a molecular connectivity calculation. Examples illustrate the method.

KEY WORDS: molecular connectivity; data-base search; molecular fragments; pharmacophores; toxicophores.

INTRODUCTION

Substructure or fragment searching through large compound data bases has become a prominent procedure in industrial drug design, government lab chemical hazard study, and academic research. Preliminary studies or predictions identify a critical molecular fragment which is part of a pharmacophore, active site, or toxicophore. Some scheme must be employed to identify molecules in a large data base containing that critical fragment.

Substructure searching is based upon examination of machine-readable codes in files. The most common types of codes are the linear types (1) and the well-known connection table. In substructure or fragment searching, the intent is to obtain from a large data base all molecules which contain the query substructure.

It is judged that substructure searching is more difficult than structure searching because the usual use of hash tables and canonical structure schemes is ineffective. As Willett states, "It is not possible to associate a specific and unique identifying code with a query substructure which could then be used to identify those compounds that contained it" (2). The usual strategy involves the use of a number of non-unique characteristics as screens on a data file. These characteristics are necessary but not sufficient to guarantee unique identification.

The typical substructure searching strategy is carried out in a two-step process. A screen search is first carried out to obtain the set of molecules which might contain the query substructure. In the second step the presence of the substructure is determined by a detailed atom-by-atom matching procedure. Such a process is a subgraph isomorphism problem. Such procedures are very time-consuming (3) and most efforts in substructure searching are being focused on the screening methods (4,5).

One approach to structure searching is the use of a canonical form of the connection table such as proposed by Morgan (6). A particular application of this scheme leads to the SEMA name (7,8). Mauer and Lewis (9) introduced a hashing function to calculate a code from the connection table. Such systems have been described by Bawden *et al.* (10) and by Freeland *et al.* (11).

When searching is performed using substructure codes, the problem becomes more complex and time-consuming (12,13). Craig and Ebert (14) developed manually derived fragment codes. Various fragment codes have been developed from connection tables (15). Adamson *et al.* (16) included atom-centered fragments which are similar to the work of Feldmann (17).

The basic problem is to find an index which uniquely or characteristically encodes the structure of a molecular group or fragment. The index must be relatively independent of the environment since that group or fragment will be found embedded in a molecule data base. The index must be simply calculated, unambiguous, and capable of being modified to reflect different structural environments. An excellent candidate for such an index arises from differential molecular connectivity indexes.

DIFFERENTIAL MOLECULAR CONNECTIVITY INDEXES

In earlier studies, we have shown that a significant amount of information is resident in the numerical value of the difference between the simple and the valence molecular connectivity indexes of several orders, *m*, ${}^{m}X - {}^{m}X^{v}$ (18–20). We can refer to this differential as $\Delta^{m}X$. The derivation,

[1] Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond, Virginia 23298.

[2] Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170.

[3] To whom correspondence should be addressed.

significance, and use of molecular connectivity indexes are well documented and a program is available which facilitates the calculation (18).

The $\Delta^m X$ value for a fragment is essentially a value associated with the non-Csp^3 atoms of that fragment. To illustrate, consider the ethylamine molecule (fragment):

$$CH_3 - CH_2 - NH_2 \rightarrow \wedge\!\!\!\searrow NH_2$$

If we reduce the molecule to the hydrogen-suppressed graph and then calculate the first-order simple and valence connectivity subgraphs, we see the results in Table I.

The $\Delta^1 X$ value is ${}^1X - {}^1X^v = 0.299$. This index reflects the electronic structure of the nitrogen and its immediate environment in just the alpha position. The electronic identity of the $-NH^2$ resides in the δ^v value for this heteroatom:

$$\delta^v = (Z^v - h)/(Z - Z^v - 1)$$

where Z^v are the valence electrons, Z is the total number of electrons, and h is the count of hydrogen atoms on the heteroatom.

Note that the differential index $\Delta^1 X$ excludes the $CH_3 - CH_2 -$ subgraph from the numerical result since the δ and δ^v values for these two atoms (groups) are identical. We can conclude that the index $\Delta^1 X = 0.299$ is a characteristic index for all molecular fragments in which a primary amine is bonded to one CH_2 group. This characteristic makes it possible to identify this fragment in any molecule in any structural circumstance.

Using this same reasoning, we may characterize numerically the ethylamine fragment $-CH_2-CH_2-NH_2$ by invoking the second-order information resident in $\Delta^2 X = {}^2X - {}^2X^v$. This value arises from calculations of the simple and valence connectivity indexes of the fragment in question. In this case ${}^2X = (2 \cdot 2 \cdot 1)^{-0.5}$ and ${}^2X^v = (2 \cdot 2 \cdot 3)^{-0.5}$. The differential index $\Delta^2 X = 0.500 - 0.289 = 0.211$ is characteristic of the ethylamine fragment.

We can now say that any molecule in a data base which has a second-order subgraph differential of $\Delta^2 X = 0.211$ is a candidate molecule for the presence of an ethylamine fragment.

If the fragment is $-CH_2-CH_2-CH_2-$, then the $\Delta^2 X$ value is zero and the presence of this fragment in a molecule would not be recognized. The likelihood of there being an interest in such a fragment is small since heteroatoms and non-Csp^3 atoms are the focus of attention in biological problems. We are usually interested in such a fragment when it is appended to a heteroatom or when it bridges two heteroatoms which we deem are significant in drug design.

Table I. Information in the ${}^1X - {}^1X^v$ Index

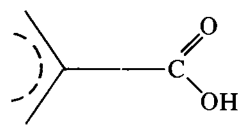
Subgraph	Simple index, $(\delta_i\ \delta_j)^{-0.5}$	Valence index, $(\delta^v_i\ \delta^v_j)^{-0.5}$
–	0.707	0.707
$-NH_2$	0.707	0.408
Sum of subgraph indexes	${}^1x = 1.414$	${}^1X^v = 1.115$

MOLECULAR-GROUP IDENTITIES

Using this same concept, we can characterize any molecular group or fragment with an index of some appropriate order and then use this index to search for candidate molecules possessing such a fragment.

In Table II the characteristic indexes for some common functional groups which might be targets of data-base searches are calculated. We cannot say at this time that any characteristic index is uniquely associated with one group. Our feeling based on extensive experience is that redundancies would be very rare. If such a situation were to arise, then unique descriptions would be found using other models of the fragment leading to other orders of $\Delta^m X$.

Further refinements in the description of the groups in Table II are possible guided by the particular requirements of a data-base search. Suppose, for example, that we wished to encode and then search for carboxyl groups attached to aromatic rings rather than aliphatic structures. The aromatic carboxyl group would be modeled to include the ipso atom of the ring:



Using order 3 cluster indexes, $\Delta^3 X_c = 0.333$, $\Delta^3 X^v = 0.046$, and $\Delta^3 X_c = 0.288$. The corresponding index for aliphatic acid groups $-CH_2-COOH$ is $\Delta^3 X_c = 0.344$.

Again referring to the examples in Table II, if we wished to encode, then search selectively for, different classes of amines, we could appropriately model the fragments to include the desired environment. For aliphatic primary amines, $-CH_2-NH_2$, $\Delta^1 X = 0.299$; for aliphatic secondary amines, $-CH_2-NH-CH_2-$, $\Delta^2 X = 0.104$. The latter case is distinguished from $-CH_2-CH_2-NH_2$, which has $\Delta^2 X = 0.211$. Finally, a tertiary amine in the environment $(-CH_2)_3-N$ is characterized by the third-order cluster differential index $\Delta^3 X_c = 0.046$.

Using the same approach, we can characterize various environments of functional groups, calculate the characteristic differential connectivity index, and then use it in a search of a data base for the presence of that fragment.

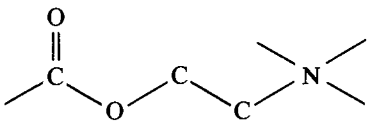
Table II. Characteristic $\Delta^m x$ Indexes of Several Common Functional Groups

Group (fragment) structure	Order, m, of index	Value of $\Delta^m x$
$-COOH$	2	0.616
$-CO-O-$	2	0.624
$-NO_2$	2	0.633
$-CN$	1	0.483
$-Ph$	5	0.070
Aryl-Cl	1	0.010
Aryl-Br	1	−0.405
$-N(CH_2-)_2$	2	0.065
$-CH_2OH$	1	0.391

LARGE-FRAGMENT IDENTITIES

Beyond a search for functional groups, it is possible to search for larger fragments of interest. Preliminary research may identify a molecular fragment believed to be critical to the action of an agonist or antagonist, to a substrate or inhibitor, or to some pharmacodynamic property.

As an example let us assume that the essential molecular fragment for cholinergic activity has been identified as



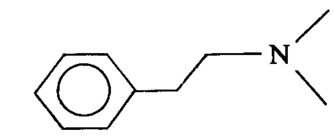
To search a data base for these atoms in this arrangement we must begin with the highest order which embraces all of the atoms in the fragment. Here it would be the fifth-order path differential connectivity index. Specifically $\Delta^5X_p = 0.083$. We can be certain that wherever this value occurs among the subgraph indexes in molecules in a data base, then we will capture those molecules containing this fragment.

The question arises, however, as to whether we may capture other fragments composed of the same atoms but in a different order. Specifically will we capture molecules containing the fragment with the ether oxygen and methylene carbon interchanged with a search index of $\Delta^5X = 0.083$?

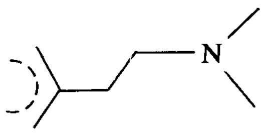
The answer is yes. This value of Δ^5X will pick up "noise" consisting of the same atoms in a different order.

How can we refine the search to eliminate these fragment isomers and capture just the desired fragment? The answer lies in a second-echelon search within just the fragments identified by Δ^5X. Table III shows the Δ^mX values calculated for subfragments of the two fragments noted above. Within the fragments isolated by the descriptor $\Delta^5X_p = 0.083$, those fragments containing $\Delta^2X = 0.325$ are discriminated from the others as being the fragments sought for. Alternatively, we can exclude the unwanted fragment by rejecting those containing $\Delta^2X = 0.264$. Several approaches to this issue are equally useful. The choice of models and their selection and rejection depend upon the problem being pursued.

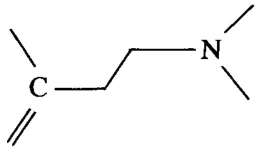
A second example illustrates the versatility of the approach. Consider searching for this phenethyl t-amine fragment in a large molecular data base.



A simple approach to modeling this fragment is to adopt the fragment



as the search structure. In this case $\Delta^3X_p = 0.055$. The search will reveal molecules containing phenethylamine fragments with all possibilities of phenyl-ring substitution. The search will also find all fragments



as unwanted "noise."

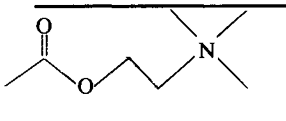
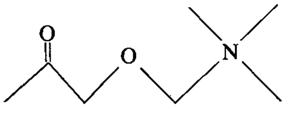
If the search is based upon a model in which the complete ring is a part, then this fragment will be discovered with $\Delta^8X_p = 0.022$. This model, however, will not reveal phenethylamine fragments in which one or two more ring positions are substituted.

Ring-substituted fragments of this class may be isolated using the appropriate model and calculating the Δ^8X_p value for the search. Specifically for monosubstituted $\Delta^8X = 0.018$, for disubstituted $\Delta^8X = 0.014$, for trisubstituted $\Delta^8X = 0.011$, and so forth. A further refinement in the search can reveal desired positional arrangements of ring substitution by using a second-echelon search among fragments, searching for ortho, meta, para, etc., relationships.

DATA CONFIGURATION

It is obvious that the data base must be configured for searches based upon molecular connectivity indexes of various orders, simple and valence weighted. Each molecule is entered as a connection matrix with the valence weights in-

Table III. Δ^mX Profiles for a Fragment

Fragment	Δ^5X_p	Δ^4X_p	Δ^3X_p	Δ^2X_p
	0.083	0.162	0.230	0.325
		0.056	0.102	0.144
			0.085	0.149
				0.026
	0.083	0.162	0.230	0.264
	0.056	0.102	0.144	
			0.085	0.149
				0.121

cluded. This is the standard procedure for instituting a molecular connectivity analysis of a molecule.

From this input, the simple and valence-weighted indexes of all orders and types are calculated and made accessible to a search procedure. The program MOLCONN written by Hall[4] has been developed with this use in mind.

After the decision is made as to how to model the fragment, the appropriate $\Delta^{m}X$ index is fed to the data base. If second-echelon searching is required, each candidate fragment in the candidate molecules is treated as a complete molecule and subgraphs within are searched to discriminate among connectivity isomers.

In calculating the $\Delta^{m}X$ value for a modeled fragment, it is advisable to use five decimal places to obviate any possible redundancies at a lower level of precision. At this level, only structural isomerism would produce redundant values of $\Delta^{m}X$.

This general method affords a quick, reliable method to search for any fragment in a molecular data base.

REFERENCES

1. J. E. Rush. *J. Chem. Inform. Comput. Sci.* **16**:202–210 (1976).
2. P. Willett. Research Studies Press, John Wiley & Sons, Letchworth, England, 1987.
3. E. H. Sussenguth. *J. Chem. Docum.* **5**:36–43 (1965).
4. A. Feldman and L. Hodes. *J. Chem. Inform. Comput. Sci.* **15**:147–152 (1975).
5. P. G. Dittmar, N. A. Farmer, W. Fisanick, R. C. Haines, and J. Mockus. *J. Chem. Inform. Comput. Sci.* **23**:93–102 (1983).
6. H. L. Morgan. *J. Chem. Docum.* **5**:107–113 (1965).
7. A. E. Petrarca, M. F. Lynch, and J. E. Rush. *J. Chem. Docum.* **7**:154–165 (1967).
8. W. T. Wipke and T. M. Dyott. *J. Am. Chem. Soc.* **96**:4825–4834 (1974).
9. W. D. Maurer and T. G. Lewis. *Comput. Surv.* **7**:5–19 (1975).
10. D. Bawden, J. T. Catlow, T. K. Devon, J. M. Dalton, M. F. Lynch, and P. Willett. *J. Chem. Inform. Comput. Sci.* **21**:83–86 (1981).
11. R. G. Freeland, S. A. Funk, L. J. O'Korn, and G. A. Wilson. *J. Chem. Inform. Comput. Sci.* **19**:94–97 (1979).
12. J. Figueras. *J. Chem. Docum.* **12**:237–244 (1972).
13. A. von Scholley. *J. Chem. Inform. Comput. Sci.* **25**:235–241 (1985).
14. P. N. Craig and H. M. Ebert. *J. Chem. Docum.* **9**:141–146 (1969).
15. J. E. Crowe, M. F. Lynch, and W. G. Town. *J. Chem. Soc. (C)* 990–996 (1970).
16. G. W. Adamson, M. F. Lynch, and W. G. Town. *J. Chem. Soc. (C)* 3702–3706 (1971).
17. R. J. Feldmann. *Annu. Rev. Biophys. Bioeng.* **5**:477–510 (1976).
18. L. B. Kier and L. H. Hall. *Molecular Connectivity in Structure-Activity Analysis,* John Wiley & Sons, New York, 1986.
19. L. B. Kier and L. H. Hall. *Quant. Struct.-Act. Relat.* **2**:163 (1983).
20. L. H. Hall and L. B. Kier. *J. Mol. Struct.* **134**:309 (1986).

[4] The program MOLCONN may be purchased from Lowell H. Hall, Eastern Nazarene College, Quincy, Massachusetts 02170.