Structure-Activity Relationship Oriented Languages for Chemical Structure Representation

V. V. AVIDON and I. A. POMERANTSEV

Scientific Research Institute for Biological Testing of Chemicals, Moscow Region, USSR

V. E. GOLENDER and A. B. ROZENBLIT*

Institute of Organic Synthesis, Latvian SSR Academy of Sciences, Riga 226006, USSR

Received October 7, 1981

A family of chemical structure description languages for structure–activity studies is proposed. It is based on the description of mutual displacement of potentially active (descriptor) centers in molecules and includes the fragmental notation SSFN, the topological system DCAM, and the topographical system DCGM. An approach to the automated selection of pharmacophores using the above languages is discussed.

INTRODUCTION

Despite the fact that a variety of notations, systems, and languages for the representation of chemical compounds have been developed to date, new approaches are continuously emerging. This situation can be explained by the interaction of the following factors:

(1) "Chemical structure" is an exceedingly complex entity. It can be seen as (a) a set of atoms specifically connected within a molecule as represented by the structural formula indicating the presence and topology of particular functional groups or as (b) a combination of physicochemical properties complementary to the structural formula, for instance, distribution of electron density in space.

(2) The choice between different methods of structural representation is determined by (a) the ultimate aim set for structure description procedure, since during the performance of various tasks, the main emphasis is placed on different aspects of the complex concept of "chemical structure", and by (b) the level of technical means available for structure manipulations.

The aforesaid has been used as a starting point for developing the present approach to chemical structure description specifically designed for the analysis of structure–activity relationships in a structurally heterogeneous set of compounds. This approach is consistent with the current level of computer technology and the latest developments in chemical information processing.

MAIN PREREQUISITES

According to Ariens,[1] the action of bioactive compounds can be divided into stages: pharmaceutical, pharmacokinetic and pharmacodynamic. The specificity of action of a medicine is determined, first and foremost, by the pharmacodynamic stage, the former two stages being crucial for the intensity of effect.

The pharmacokinetic stage is determined by such structural characteristics as partition coefficient and electronic and steric factors which are employed in the Hansch approach and in other lead optimization methods.[2] Essentially different structural characteristics may appear crucial during the pharmacodymamic stage. These require other approaches to structure representation for the prediction of types of biological activity or lead generation.

The principles underlying the present approach to chemical structure description are based on the existing knowledge of the processes underway during the pharmacodynamic stage.

(1) The ability of a chemical compound to exert pharmacological action is determined by its capacity to bind complementarily and interact with the functional chemical structure in the living organism, i.e., the "receptor".

(2) Interaction of chemical compounds with the receptor involves only part of the molecule, the pharmacophore. This process is accompanied by the formation of weak chemical bonds: hydrogen, electrostatic, van der Waals, hydrophobic, and ionic.

(3) Compounds carrying identical or similar pharmacophores are capable of exerting identical specific effects.

On the basis of these theoretical assumptions, one can set down the following requirements for structure description languages for handling of lead generation problems. The language should provide a basis for (a) adequate representation of pharmacophores, (b) the design of efficient procedures for the identification of pharmacophores during the processing of structure and activity data obtained for chemical compounds, and (c) the development of reliable algorithms for the prediction of biological activity of compounds and lead generation based on the found pharmacophores.

At present, topographical description of pharmacophores appears to be the most adequate approach. This description makes use of interatomic distances and charge values estimated by using the results of quantum-chemical and conformational calculations. This approach was first put forward by Kier[3] and further developed by other investigators, e.g., ref 4 and 5. However, it has not become popular for the analysis of large structure–activity data files because of the complexity of calculation procedures involved. Moreover, the first attempts of biological activity prediction were commonly conducted with the aid of structural notation languages that would take no account of the specificity of tasks being solved. To this end were used fragmental notations developed for the purposes of information retrieval,[6,7] augmented atomic fragments including a central atom and its immediate environment,[8,9] parameters indicating the presence or the occurrence number of atoms or substructures of a particular type.[10,11] Further, attempts were made to include in structure representations the specific fragments describing the structural characteristics essential for biological activity. For instance, Chu et al.[12] described the use of heteropath fragments including a pair of heteroatoms and a chain of atoms connecting them; Jurs et al.[13] considered geometrical descriptors obtained by using the results of conformational calculations. Electron-density characteristics over certain types of fragments were employed too.[14]

It must be pointed out that the application of pattern recognition methods to the analysis of structure–activity relationships (see reviews[15,16]) described in the publications

0095-2338/82/1622-0207$01.25/0

mentioned above and in some other ones does not involve automatic pharmacophore detection. For example, one paper[17] describes pharmacophore selection by manual structural formula superposition conducted by the investigator prior to SAR analysis.

This paper is concerned with the analysis of structure-representation languages useful for pharmacophore description and for the design of algorithms required for their automatic detection. The languages under discussion vary both with respect to the degree of adequacy of pharmacophore descriptions and their complexity.

SUBSTRUCTURE SUPERPOSITION FRAGMENTAL NOTATION (SSFN)

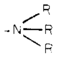
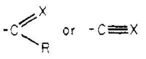
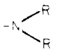
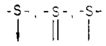
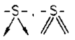
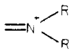
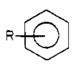
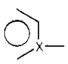
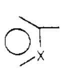
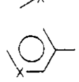
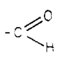
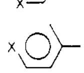
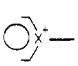
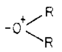
SSFN[18,19] is the simplest structure-representation language among those touched upon in this paper.

As judged by the principles of language design outlined earlier, it appears advisable during chemical structure analysis to select potentially active centers, i.e., atoms or groups of atoms functioning as centers of van der Waals type interaction or reaction centers. These centers, further referred to as "descriptor centers" (DC), first of all include atoms and groups of atoms containing valence p and d electrons or an integer electrostatic charge. Distances between DC are essential for compound–receptor interaction. The SSFN language is based on the identification of DC and distances between them.

The coding procedure involves listing of all chains in the compound initiated and terminated by DC and not passing through another DC. The chains may pass through cyclic systems too, if the latter are not aromatic. Linear descriptors describe all such chains. For a more detailed representation of structure cyclic descriptors describing the cyclic moiety are used.

Descriptor Centers. All heteroatoms, including N, O, S, P, haloid, metal, etc., are assumed as descriptor centers in the SSFN language, as well as cyclic aromatic systems as a whole and pairs of carbon atoms connected by multiple bonds (double or triple but no aromatic). DC of the SSFN language along with their codes are listed in Table I.

Table I. Descriptor Centers of the SSFN Language

DC[a]	code	DC[a]	code	DC[a]	code
	01	-SH	21	$-CH_3$, $=CH_2$, $\equiv CH$	41
$-NH_2$ or $-NHR$	02	-SR	22		44[b]
	03		23	Met (metal)	43
=NH	04		24	P	47
=NR	05	=S	25	As, Sb	51
=N	06	Cl, Br, I	31	Si, Ge	52
	07	F	32	Se	54
$-N^{-}-$	10		33	B	53
-OH	11		34	$(-C\equiv C-)_n$	45
-OR	12		35	$(-C=C-)_n$	46
=O	13		36		
	14		37		
$-O^{-}$	15		40		
	16				

[a] R = any atom except H and X = a heteroatom. [b] DC 44 is used for the description of substituents in aromatic systems.

Clearly, the electronic pattern of heteroatoms and consequently their participation in the formation of "weak" bonds is strongly affected by their valence state and the types of atoms connected with them. For this reason, the same heteroatoms may be regarded as different DC according to their environment. For instance, the quaternary nitrogen carrying an integer positive charge (DC 01) differs drastically from tertiary nitrogen (DC 03) which, in its turn, is different from secondary or primary nitrogen (DC 02) in that it is incapable of acting as a proton donor during hydrogen bonding. Conversely, primary and secondary nitrogens can belong to the same DC since they both can serve as proton donors. It is apparent that an oxygen atom attached by a double bond (DC 13) differs significantly from the same atom having two single bonds (DC 12 and 11). The latter DC strongly differ from one another because the OH group (DC 11), being a proton donor, can undergo ionization and participate in hydrogen bonding, but ester oxygen (DC 12) fails to do so.

On the other hand, atoms of various elements can belong to the same DC. For example, the effects exerted by Cl, Br, and I (DC 31) on biological activity differ quantitatively rather than qualitatively, the fluorine atom (DC 32) being different because it is capable of forming a fairly strong hydrogen bond. Some atoms belonging to the same group in the periodic system, e.g., As, Sb, Bi (DC 51) or Si, and Ge (DC 52), are attributable to the same DC not only for similarity considerations but also because they seldom occur in molecules of biologically active agents. If necessary, further development of the language may require their further transfer from the present DC to some new ones.

Terminal CH_3 groups are also regarded as DC. This permits the characterization of various lipophilic substituents playing an important part in the formation of lipophilic bonds.

Linear Descriptors. As steric-oriented stabilization of a compound on the receptor requires at least two DC in the molecule, chains of carbon atoms initiated and terminated by DC are selected as descriptor words in the SSFN language.

It is also important to indicate whether conjugation between all the atoms in the chain is feasible. Hence, SSFN descriptor is designed according to the scheme:

descriptor center 1	chain length (no. of carbon atoms)	descriptor center 2	digit indicating presence or absence of conjugation

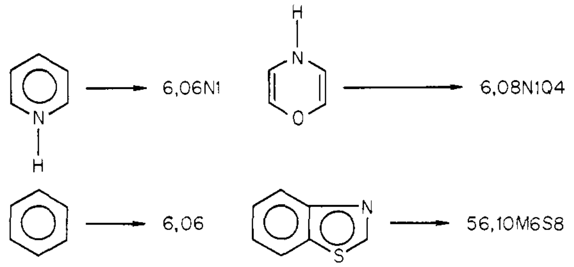
A linear descriptor is coded by a seven-digit number. Two digits are assigned to the descriptor center and chain length, respectively. Conjugation is described by a single digit, being equal to 1 when there is conjugation and equal to 0 in the absence of it. Conjugation is considered to be present if the chain length is 00 or if all carbon atoms in the chain are in sp^2 or sp hybridization.

Cyclic Descriptors. Cyclic systems are coded in the SSFN by means of cyclic descriptors[19] having the form:

"head"	"body"	"tail"
geometry of cyclic system	no. of π electrons in cyclically conjugated system	location of heteroatoms

"Head" indicates the cycle size (number of atoms) in the case of monocycles or the size and location of individual cycles in polycyclic systems. Two digits are assigned to the "body".

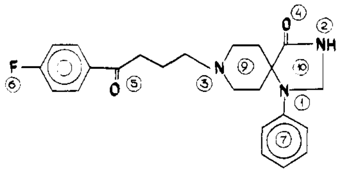
If at least one atom is in the state of sp^3 hybridization, the descriptor body is designated "00". The tail includes heteroatoms and their position in the cycle. For instance



SSFN Coding of Compounds. The codes of compounds include all linear and all cyclic descriptors listed in lexicographical sequence.

To exemplify the point, we will describe the coding of two compounds. Descriptor centers and cyclic systems of these compounds are indicated on the structural formulas by numbers.

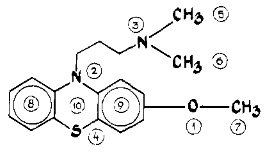
Compound I belongs to heterocyclic butyrophenones.



I

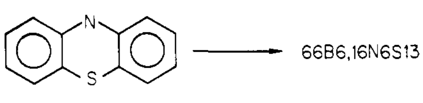
During the coding process structure I is partitioned into substructures presented along with their codes in Figure 1. The SSFN code of compound I is as follows: 5,00N1N3. 6,00N1. 6,06; 0201030. 0201131. 0202030. 0204030. 0300331. 0302130. 0303030. 0304130. 0304330. 1301331. 3200331.

Compound II belongs to phenothiazine series.

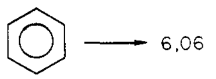


II

The cyclic moiety is coded as follows: 8–10 are coded as a single system.



Aromatic fragments eight and nine in the cyclic system are coded separately



Linear descriptors are coded as in compound I. Finally, the code of compound II is obtained: 6,06. 66B6,16N6S13; 0300331. 0301410. 0303030. 1200331. 1201410. 2200331.

Advantages and Shortcomings of SSFN. The main advantage of SSFN resides in the simplicity of representation of structural characteristics essential for the manifestation of biological activity by means of descriptor sequence. Hence, it can be simply used in statistical pattern-recognition algorithms and for quantitative evaluation of structural similarity among biologically active agents.[19,20]

SSFN language can be easily acquired by chemists. The coding of a compound requires, on the average, 2 min, its code containing about 14–15 descriptors.

The main disadvantage of SSFN and other fragmental notation systems is the "disintegration" of structure. The information stored concerns the structural fragments of the molecule only but fails to provide information on their interconnections. Moreover, manifestation of biological activity frequently depends on the overall distance between two active centers which can be separated not only by a carbon chain but also by heteroatoms. This chain in SSFN is broken into several descriptors and cannot be found as an integral moiety.

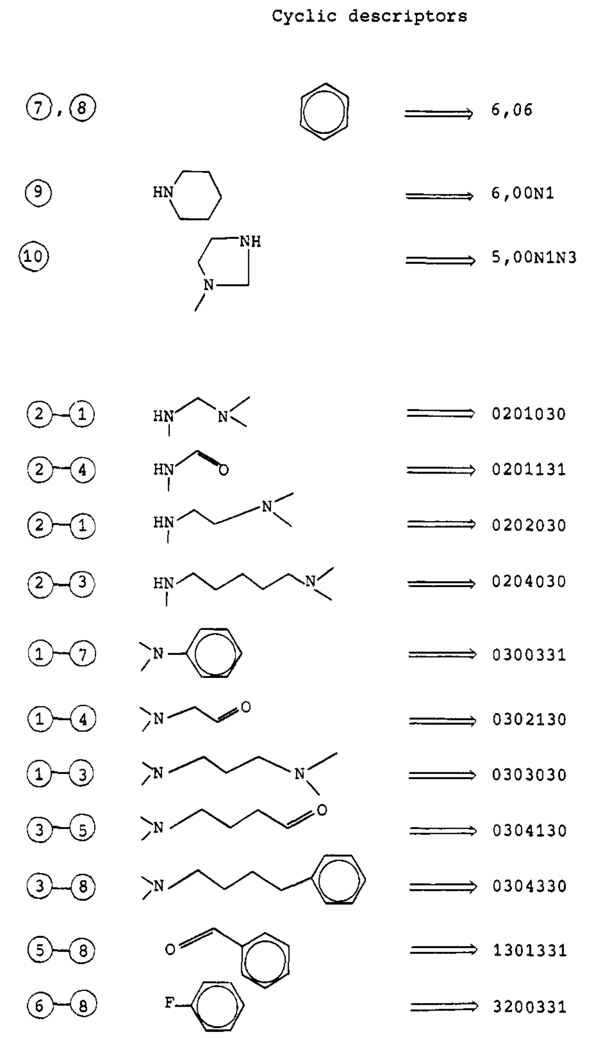


Figure 1. SSFN descriptors for compound I and corresponding substructures.

Further development of the main principles of the SSFN language has resulted in a new method of structural representation which, satisfying the chief requirements of this language, retains the topology of a molecule as a whole. The new form of the language will be referred to as a descriptor centers adjacency matrix (DCAM).

TOPOLOGICAL SYSTEM DCAM

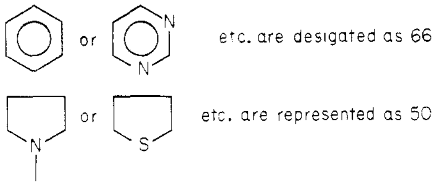
The main principles of the DCAM topological system are as follows: potentially active centers or descriptor centers (DC) in a compound are selected and coded; the DCAM system contains information concerning the length and character of atom–atom chains between all descriptor centers.

In keeping with these principles, the structure of a compound can be represented by an undirected graph, its vertices corresponding to descriptor centers and its edges showing the paths between descriptor centers. Vertex types correspond to types of descriptor centers and edge types to path length (number of atom–atom bonds). The edge-type label may also contain information on the nature of chemical bonds along the path. This graph will be denoted as a descriptor center connection graph (DCCG). Accordingly, the adjacency matrix of this graph will be designed as descriptor center adjacency matrix (DCAM).

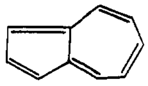
Interpretation of Descriptor Centers for DCAM. As the DCAM language is a more elaborate version of the SSFN language, the meaning of DC remains unchanged. Descriptor centers in DCAM, like in SSFN, are represented by heteroatoms (see Table II), but unlike the case of SSFN, they also include heteroatoms of aromatic cycles.

DCAM lacks special descriptors for cyclic systems; therefore, to gain necessary information concerning cycles, we considered any single cycle not exceeding nine atoms as a descriptor center. The code of such a descriptor center consists of two digits, the first digit indicating cycle size and the second showing the number of π electrons in the conjugated cyclic system. If any of the atoms in the cycle lacks π electrons (conjugated cyclic system is absent), zero is indicated as the second digit.

For example, consider the following cyclic systems:

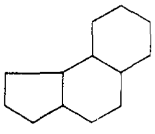


In the cases when a single cycle is not strictly aromatic, i.e., when the number of π electrons in its conjugated cyclic system fails to satisfy the Hückel rule ($4N + 2$; where N is an integer) but is a part of a fused cyclic system, which is on the whole aromatic, the second number is set equal to 6 (which is in compliance with the Hückel rule). This approach gives information indicating that a given cycle consists of aromatic bonds only. For example, the azulene molecule contains five

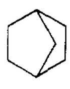


π electrons in the five-membered cycle and seven π electrons in the seven-membered one; i.e., strictly speaking, each of the two cycles is not aromatic. Nevertheless, since azulene, as a whole, is aromatic (ten π electrons) the cycles constituting it will be designated "56" and "76", respectively.

All cycles constituting the minimal basis, as well as any other cycles containing less than ten atoms, are considered as DC. Examples are represented by polycyclic systems III and IV. For the fused system III, three cyclic DC are indicated



III

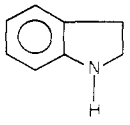


IV

as 50, 60, and 60. For bridged system IV three DC are also indicated: 50, 50 (basic cycles), as well as 60.

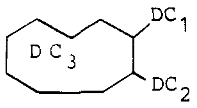
Nondiagonal Entries of DCAM. The nondiagonal entries of DCAM correspond to the edges of DCCG graph and, consequently, must contain information on atomic chains connecting descriptor centers.

Pharmacological data suggest that manifestation of biological activity does not depend exclusively on the shortest chains connecting the active centers. For instance, the chain consisting of three bonds and connecting the nitrogen atom with the benzene ring in the indoline structure shown below



is more important for the manifestation of neuroleptic activity than the shortest chain consisting of one bond only, thus indicating a similarity between indoline and phenylethylamine.

It is necessary, therefore, to indicate in DCAM not only the shortest possible chain but also several chains connecting the ith and the jth descriptor centers. In order to avoid listing all such chains which can be too numerous in complex structures, it is reasonable to impose some restrictions. The shortest and other possible chains not exceeding ten bonds, therein including chains equal in length, should be indicated. This permits one to retain information on the second chain connecting DC_1 and DC_2 in the structures like



containing the biggest possible (nine-membered) cycle regarded as an individual DC.

Description of nondiagonal matrix elements also includes, besides chain length (number of bonds), information on the presence or absence of conjugation between DC along this chain. Let us assume that each edge is characterized by a three-digit number of which the first two digits designate the number of interatomic bonds in the chain; the last one carries information on the type of bonds in the path and the possibility of conjugation along it. The formal feature of conjugation is present when all carbon atoms in the chain connecting the ith and the jth DC are in an sp^2 hybridization state (this applies equally to the key carbon of the carbon DC, i.e., the atom to which the chain in question is attached).

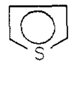
The third digit in the edge label has the following meanings: 0,1, all bonds in the chain are single; 2,3, multiple bonds are present; 4,5, aromatic bonds are present (along with other ones); 7, all bonds are aromatic.

An even number indicates that conjugation along the chain is absent, but an odd number shows that conjugation is possible.

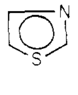
Mutual Overlapping of Descriptor Centers. DC consisting of several atoms exist in the DCAM topological system. They include all cyclic DC as well as pairs of carbon atoms connected by multiple bonds (not included in aromatic cycles). DC containing one or several common atoms will be referred to as mutually overlapping DC. In this case the DCCG graph edge connecting the two DC (and the appropriate DCAM element) will be labeled by a value denoting the number of common atoms in the given DC with a minus sign. For example, in the pyridine structure, the following DC are selected: nitrogen ("05") and cycle ("66"). These DC are considered as mutually overlapping, and the corresponding DCAM entry is −1. Two cycles sharing a common edge will have the overlap label −2, whereas the union of two spirocycles will be labeled −1. Overlap indices allow one to retain information on the structure of polycyclic systems.

The complete version of the DCAM language is rather cumbersome; therefore, for practical reasons it is advisable sometimes to apply its abridged versions containing information on the shortest distances between DC only, the rest of the information being omitted (bond characteristics and so on).

Examples of Structure Representation Using the DCAM Language. The first example will demonstrate the procedure of common subgraph selection performed for two heterocycles, thiophene and thiazole, using DCAM topological system.



$$\begin{vmatrix} \underline{22} & -1 \\ & \underline{56} \end{vmatrix}$$



$$\begin{vmatrix} \underline{22} & -1 & 027/037 \\ & \underline{56} & -1 \\ & & \underline{05} \end{vmatrix}$$

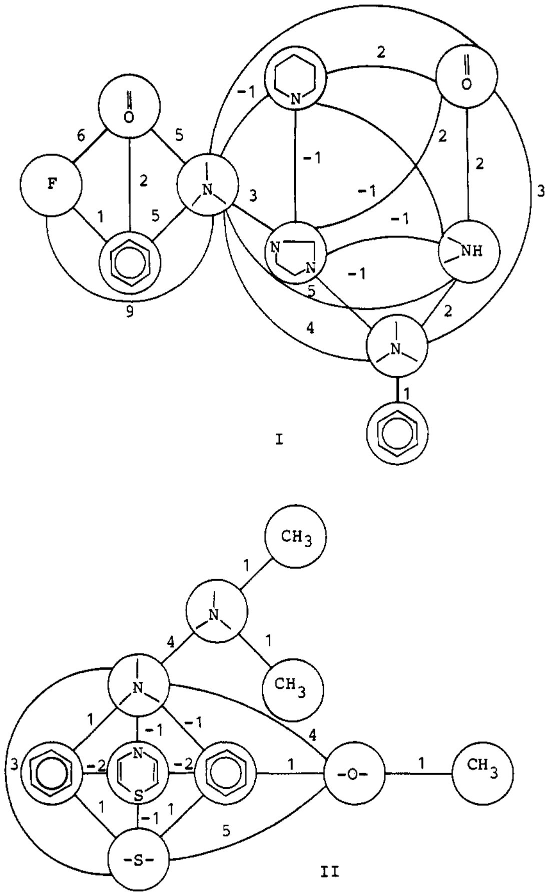
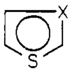
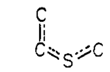


Figure 2. Descriptior center connection graphs for compounds I and II.

As can be seen, the thiophene DCAM is a part of the thiazole DCAM. This corresponds to the common fragment

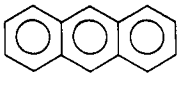


If thiazole and thiophene are represented as atom adjacency matrices, the common fragment consisting of four atoms only can be identified.

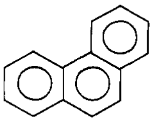


Therefore, owing to additional meaningful information, the results of the DCAM comparison are more consistent with chemical intuition than those obtained by using adjacency matrices.

As can be seen below, DCAM is also useful for the selection of similarity/dissimilarity features in such tricyclic systems as anthracene and phenanthrene.



$$\left|\begin{array}{ccc} \underline{66} & -2 & 027/027 \\ & \underline{66} & -2 \\ & & \underline{66} \end{array}\right|$$



$$\left|\begin{array}{ccc} \underline{66} & -2 & 017/037 \\ & \underline{66} & -2 \\ & & \underline{66} \end{array}\right|$$

Comparison of the DCAM for these compounds leads to the conclusion that they both consist of three six-membered cycles, each pair of them sharing a common edge. At the same time, variation in matrix elements 1,3 allows one to distinguish between these structures.

	1	2	3	4	5	6	7	8	9	10
1	03	020	040	032	092	134	011	090	010	-1
2		02	050	023	102	144	030	100	020	-1
3			03	052	052	094	050	050	-1	030
4				13	102	144	042	102	022	013
5					13	065	102	023	052	082
6						32	144	011	094	124
7							66	100	020	011
8								66	050	080
9									60	-1
10										60

Figure 3. DCAM for compound I.

	1	2	3	4	5	6	7	8	9	10
1	12	045	084	055	094	094	010	055	011	035
2		03	040	035	050	050	054	011	011	-1
3			03	074	010	010	094	050	050	040
4				22	084	084	064	011	011	-1
5					41	020	104	060	060	050
6						41	104	060	060	050
7							41	064	020	044
8								66	021	-2
9									66	-2
10										68

Figure 4. DCAM for compound II.

The DCAM representation of more complex compounds will be exemplified by the DCAM of compounds I and II. Figure 2 shows the DCCG of these compounds (for the sake of clarity only edges corresponding to bonds between adjacent descriptor centers are indicated): the DCCM of these compounds are presented in Figures 3 and 4 (only the shortest paths).

DCAM contains ample information on the presence of active centers in the molecule and their mutual connections. This information, however, is restricted to molecular topology and provides only a qualitative reference to the electronic structure of some atomic groups, despite the fact that the spatial and electronic structure of chemical compounds is crucial for the manifestation of their biological activity and other properties.

It is possible to further develop the matrix representation of chemical structures described by DCAM. The first step toward this end is achieved by the introduction of a descriptor center geometry matrix (DCGM).

TOPOGRAPHICAL SYSTEM DCGM

The DCGM system retains the same definitions of descriptor centers as those used in DCCM.

Geometric distances between descriptor centers are used as nondiagonal matrix elements of DCGM. During the determinations of distances between DC, coordinates of DC containing more that one atom are found by averaging the coordinate values of the contributory atoms.

For cyclic DC, besides the distances between them, angles describing mutual spacing of centers can be introduced: (a) the angle formed by the straight line connecting the centers and plane of the cyclic center, for point and cyclic centers; (b) angles formed by the plane and the straight line connecting

	1	2	3	4	5	6	7	8	9	10
1	03	2.25	4.28	3.53	8.26	11.23	2.81	9.47	2.90	1.20
2		02	4.97	2.27	9.25	13.07	4.93	11.00	3.61	1.16
3			03	3.66	4.80	8.97	5.79	6.46	1.42	4.16
4				13	7.98	12.50	6.24	10.08	2.72	2.44
5					13	6.21	8.58	3.46	5.93	8.39
6						32	10.11	2.98	9.82	11.94
7							66	9.04	4.78	3.98
8								66	7.53	9.95
9									60	2.74
10										50

Figure 5. DCGM of compound I.

	1	2	3	4	5	6	7	8	9	10
1	12	4.75	6.11	5.85	7.12	6.96	1.43	7.31	2.73	4.90
2		03	4.99	3.04	5.44	6.21	6.06	2.82	2.81	1.07
3			03	7.91	1.50	1.49	7.40	6.64	6.35	5.89
4				22	8.19	9.20	6.72	3.12	3.13	2.02
5					41	2.46	8.41	6.47	7.01	6.22
6						41	8.19	7.93	7.53	7.18
7							41	8.52	3.69	6.08
8								66	4.94	2.49
9									66	2.49
10										68

Figure 6. DCGM of compound II.

the centers and angles between the planes of the centers, for two cyclic descriptor centers.

For nonplanar cycles, the plane is found by the least-squares method.

Angle parameters can be omitted if they are inessential for the solution of a given problem.

The above characteristics may vary over a certain interval, in this case the range of variation is indicated.

If a molecule has several possible conformations, its description must include the DCGM for each of the conformations.

The procedure will be exemplified by using the DCGM of compounds I and II (see Figures 5 and 6). The distances between DC were calcualted by using X-ray data.[21,22]

A further development of DCGM is CCGM (charged center geometry matrix). In this case, the labels of descriptor centers are replaced by their electronic and spatial characteristics. These parameters may include, e.g., the effective charge on the atom, π-electron density, van der Waals radius, etc. Bonds may be also characterized by several parameters, e.g., bond order.

PRINCIPLES OF AUTOMATED SELECTION OF PHARMACOPHORES

The described languages can be employed to design algorithms for automated selection of structural features responsible for biological activity, i.e., pharmacophores. All the algorithms used are based on the logicostructural approach,[15,23] the cornerstone of this approach being the following statement derived from Mill's laws of inductive logic.[24]

If compounds C^1, C^2, ..., C^n, the structures of which are described in the terms of some language as S_1, S_2, ..., S_n, exert a certain biological effect A, this action may be caused by a particular structural fragment, F, incorporated in the structure of each of these compounds. The structure of this fragment can be described by the intersection of structure descriptions of all the compounds in question:

$$S_F = \bigcap_{i=1}^{n} S_i$$

In practice, the intersection of the complete set of known structures possessing a particular activity can appear empty because of the absence of a common substructure within the framework of a given structure-representation language. Such a situation may be due either to the existence of several pharmacophores responsible for a certain type of activity or to the inability, within this particular notation system, to establish structural similarity of compounds. In this case, it is necessary to intersect different subsets of the complete compound set. In other cases, although the intersection is not empty, i.e., a certain common fragment can be identified, the selected feature is statistically insignificant because of its high occurrence rate in inactive compounds. A more detailed description of the procedure for the selection of significant features (pharmacophores) was presented elsewhere.[15,23,26]

Thus, the structure-representation language designed for the selection of biological activity features must be supplemented with the definition of structural notations intersection.

For the SSFN language this intersection is based on the definition of intersection of sets; i.e., intersection of compound descriptions in the SSFN results in common descriptors for these compounds.

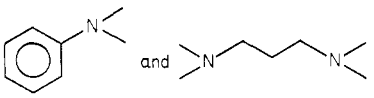
For the DCAM language and other languages based on graph representation, a definition of graph intersection is introduced, viz., the maximal subgraph common for all intersected graphs. The maximal common subgraph is a subgraph such that the addition to which of at least one vertex or edge makes it impossible for it to be a subgraph for at least one of the intersected graphs. It must be pointed out that unlike a single possible intersection of data sets, there can exist several nonisomorphic intersections of labeled graphs. Several procedures for the identification of maximal common subgraphs have been proposed.[25-27] During the execution of the intersection procedure it is possible to check not only the equality of matrix elements but also their coincidence up to a given threshold. Such a modification is especially essential when DCMG or other topographical systems are used.

To demonstrate the principles of the intersection procedure using the languages described in this paper, we will execute intersection of various structural notations of compounds I and II.

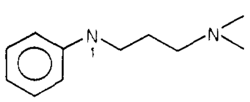
Intersection of the SSFN codes results in three descriptors:

$$6.06;\ 0300331.\ 0303030 \qquad (I_1)$$

These codes correspond to substructures



These substructures in the above compounds are combined in a single fragment:



F_1

But intersection I_1 does not formally lead to this conclusion.

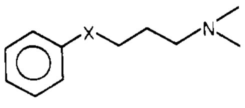
Intersection of DCAM for compounds I and II results in the following submatrices

$$\begin{vmatrix} 03 & 4 & 1 \\ & 03 & 5 \\ & & 66 \end{vmatrix} \qquad \begin{vmatrix} 03 & 5 \\ & 66 \end{vmatrix}$$

$$I_2 \qquad\qquad I_3$$

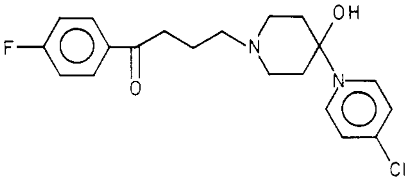
The I_2 matrix corresponds to fragment F_1.

However, the presence of nitrogen 1 is not necessary. The I_3 matrix corresponds to fragment F_2, where X stands either



F_2

for nitrogen or carbon. If we take another compound belonging to the heterocyclic butyrophenone series, haloperidol (V), a well-known pharmaceutical, then the intersection of



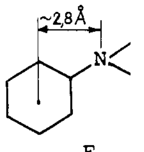
V

compounds I, II, and V will yield the P_3 matrix, i.e., fragment F_2. This fragment serves as a potent feature responsible for neuroleptic activity.[28]

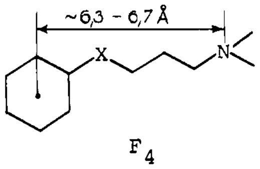
Now, we will examine the intersection of the DCGM of compounds I and II. We assume that variation in the distances between descriptor centers must not exceed 5%. Intersection of these DCGM will then give submatrices

$$\begin{vmatrix} 03 & 2.81 \\ & 66 \end{vmatrix} \qquad \begin{vmatrix} 03 & 6.64\text{-}6.35 \\ & 66 \end{vmatrix}$$

corresponding to the substructures F_3 and F_4 of the starting compounds.



F_3



F_4

In the first compound (I), substructure F_3 is obtained for descriptor centers 1 and 7, whereas in compound II it is found for descriptor centers 2 and 8 or 2 and 9.

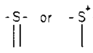
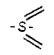
Substructure F_4 in compound I corresponds to descriptor centers 3 and 8, connected by a chain of carbon atoms, whereas in compound II they are descriptor centers 3 and 8 or 3 and 9 connected by a chain of atoms, nitrogen atom 2 being included therein. The distances between the centers amount fo 6.35 and 6.64 Å, respectively.

Interestingly, nitrogen 3 and benzene ring 7 in compound I are spaced 5.8 Å apart, and, consequently, this substructure is not included in the intersection, as opposed to the results obtained with DCAM notation. The fragment consisting of two nitrogen atoms connected by a carbon chain is also rejected in this case because in compound I the distance between nitrogen atoms (descriptor centers 1 and 3) is 4.28 Å, whereas in compound II the appropriate distance is 4.99 Å.

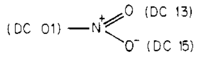
EXTENSION OF THE STRUCTURE REPRESENTATION LANGUAGES

It is not infrequent that the language chosen for chemical structure description appears inadequate for the solution of a given problem and fails to select meaningful biological activity features. In this case, it is expedient to introduce generalizations for some language elements. These generalizations must be based on the meaningful analysis of data. After the investigator has provided precise definitions of such generalizations, redescription of chemical structures must occur automatically.

Table II. Descriptor Centers of DCAM Topological System

code	structure[a]	length	code	structure	length
00	$-N^{-}-$	0	23		0
01	$-N^{+}<$ [b]	0	24		0
02	$-NH_2$ or $-(NH)R$	0	25	=S	0
03	$-NR_2$	0	31	Cl, Br, I	0
04	=NH	0	32	F	0
05	=NR	0	41	$-CH_3$	0
06	≡N	0	43	Met	0
07	$=N^{+}<$ or $\equiv N^{+}-$	0	45	-C≡C-	1
11	-OH	0	46	>C=C<	1
12	-OR	0	47	-C=C=C-	2
13	=O (except aldehyde)	0	50	P	0
14	(CH)=O (aldehyde)	0	51	As, Sb, Bi	0
15	$-O^{-}$	0	52	Si, Ge	0
21	-SH	0	54	other metalloids	0
22	-SR	0			

[a] R = any atom except H. [b] Nitrogen atom on NO_2 group is designated as DC "01":



The generalization procedure is carried out as shown below. If we assume that some notation fragments $f_1, f_2, ..., f_n$ share common properties, a new element can be introduced which is the disjunction of the above elements:

$$\varphi = f_1 V f_2 V \ldots V f_n$$

This new element φ will serve to replace elements f_i in the descriptions. The set f_i, $i = 1, ..., n$ can be described either by enumerating its elements or by applying some formal rule that enables one to associate element f_k with the set $\{f_i\}$.

Here are given some examples.

Example 1. Descriptor centers 02, 04, and 11 (Table II) can be grouped together according to their ability to act as proton donors in the formation of hydrogen bonding.

Example 2. It is assumed that the manifestation of biological activity depends on the total size of the R_1 substituent in a certain parent moiety. Then, a new element, R_v, is introduced to replace the R_1 substituent, if its van der Waals volume falls within the given range $V_{\min} < V < V_{\max}$.

The generalization procedure conducted in the interactive mode allows one to improve the language of chemical structure description by adapting it to the solution of specific problems, and provides rapid verification of hypotheses concerning the mechanism of biological action of chemical compounds.

APPLICATION OF THE PROPOSED LANGUAGES FOR SAR STUDIES

The SSFN has been used to set up a data base storing information on pharmaceuticals and biologically active compounds.[23,29,35] At the present time it stores about 9000 compounds. This data base has been used to predict biological activity of compounds by the substructural analysis method[30,31] and by applying the logicostructural approach.[23,32] These experiments have led to interesting results enabling practical implementation of prediction algorithms. The SSFN language also appeared convenient for the analysis of structural similarity between new compounds and those with known biological activity.[19,20]

The TOPLOG software package designed for the selection of topological and topographical biological activity features using the DCAM, DCGM, or CCGM languages has been developed.[33,34] The package is applicable to the selection of biological activity features either in a single large series or in a heterogeneous set of compounds.

The developed software (Fortran programs for a Hewlett-Packard HP-1000 computer system) includes special programs for structure diagram input and conversion into SSFN or DCAM codes. The input system employs an HP 2648 graphic display for structure entry using a set of commands providing creation of cyclic systems, branches, bonds, nonplanar atoms, etc. Along with a graphic image, a connection table is generated. Connection tables are used by an encoding program to detect cycles, descriptor centers, and paths between descriptor centers necessary for generation of SSFN or DCAM notations. Detailed descriptions of the algorithms used by the structure input and conversion programs will be presented elsewhere.

REFERENCES AND NOTES

(1) Ariens, E. J.; Simonis, A.-M. "Design of Bioactive Compounds". *Top. Curr. Chem.* **1974**, *52*, 1–61.

(2) Martin, Y. P. "Advances in the Methodology of Quantitative Drug Design". In "Drug Design"; Ariens, E. J., Ed.; Academic Press: New York, 1974; Vol. VIII.

(3) Kier, L. B. "Receptor Mapping Using Molecular Orbital Theory". In "Fundamental Concepts in Drug-Receptor Interactions"; Academic Press: New York, 1970.

(4) Pullman, B. "The Adventures of a Quantum-Chemist in the Kingdom of Pharmacophores". In "Molecular and Quantum Pharmacology"; Bergman, E. D., Pullman, B., Eds.; D. Reidel: Boston, 1974; pp 9–37.

(5) Kauffman, J. J.; Kerman, E. "The Structure of Phsychotropic Drugs (Including Theoretical Prediction of a New Class of Effective Neuroleptics)". *Int. J. Quantum Chem.* **1974**, 259–287.

(6) Harrison, P. J. "A Method of Cluster Analysis and Some Applications". *Appl. Stat.* **1968**, *17*, 226–236.

(7) Cramer, R. D., III; Redl, G.; Berkoff, C. E. "Substructural Analysis. A Novel Approach to the Problem of Drug Design". *J. Med. Chem.* **1974**, *17*, 533–535.

(8) Chu, K. C. "Use of Pattern Recognition and Cluster Analysis to Determine the Pharmacological Activity of Some Organic Compounds". *Anal. Chem.* **1974**, *46*, 1181–1187.

(9) Adamson, G.; Bush, J. "A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.

(10) Kowalsky, B. R.; Bender, C. F. "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test". *J. Am. Chem. Soc.* **1974**, *96*, 916–918.

(11) Stuper, A.; Jurs, P. "Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques". *J. Am. Chem. Soc.* **1975**, *97*, 182–187.

(12) Chu, K. C.; Feldman, R. J.; Shapiro, M. B.; Hazard, G. F.; Geran, G. I. "Pattern Recognition and Structure–Activity Relationship Studies. Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System". *J. Med. Chem.* **1975**, *18*, 539–545.

(13) Jurs, P. C.; Chou, J. T.; Yuan, M. "Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. A Heterogeneous Data Set"; *J. Med. Chem.* **1979**, *22*, 476–483.

(14) Chou, J. T.; Jurs, P. C. "Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. An *N*-Nitroso Compounds Set"; *J. Med. Chem.* **1979**, *22*, 792–797.

(15) Golender, V. E.; Rozenblit, A. B. "Computer-Assisted Methods of Drug Design"; Zinatne: Riga, USSR, 1978 (in Russian).

(16) Kirscher, G. L.; Kowalski, B. R. "The Application of Pattern Recognitiion to Drug Design". In "Drug Design"; Ariens, E., Ed.; Academic Press: New York, 1979; Vol. VIII, pp 73–131.

(17) Cammarata, A.; Menon, G. K. "Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores"; *J. Med. Chem.* **1976**, *19*, 739–748.

(18) Avidon, V. V.; Leksina, L. A. "Descriptor Language for the Analysis of Structural Similarity of Organic Compounds". *Nauchno.-Tekhn. Inf., Ser. 2* **1974**, *No. 3*, 22–25 (in Russian).

(19) Avidon, V. V.; Kozlova, S. P.; Arolovich, V. S. "Notation of Cyclic Fragments for the Analysis of Structural Similarity of Organic Compounds". *Nauchno.-Tekhn. Inf., Ser. 2* **1974**, *No. 12*, 21–23 (in Russian).

(20) Avidon, V. V.; Arolovich, V. S. "Analysis of Structural Similarity of Compounds Using SSFN". *Nauchno.-Tekhn. Inf., Ser. 2* **1975**, *No. 5*, 26–31 (in Russian).

(21) Koch, M. H. J. "The Crystal and Molecular Structure of 8-[3-(*p*-Fluorobenzoyl)propyl]-1-phenyl-1,3,8-triaza-spiro[4,5]decan-4-one, $C_{23}H_{26}N_3O_2F$". *Acta Crystallogr., Sect. B* **1973**, *B29*, 379–382.

(22) Marsau, P.; Gautnier, J. "Structure Cristalline de Derives de la Phenotiazine. V Methoxypromazine". *Acta Crystallogr., Sect. B* **1973**, *B29*, 992–998.

(23) Golender, V. E.; Rozenblit, A. B. "Logico-Structural Approach to Computer-Assisted Drug Design". In "Drug Design"; Ariens, E., Ed.; Academic Press: New York, 1980; Vol. IX, pp 300–337.

(24) Mill, J. S. "A System of Logic, Rationative and Inductive, Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation", 8th ed.; Harper: New York, 1900.

(25) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. M. "Molecular Structure Comparison Program for the Identification of Maximal Common Substructures". *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.

(26) Gitlina, L. S.; Golender, V. E.; Rozenblit, A. B. "Algorithms for Selection of Activity-Related Topological Features of Chemical Compounds". Abstracts of Papers of the 4th International Conference on Computers in Chemical Research and Education, Novosibirsk, USSR, 1978.

(27) Levi, G. "A Note on the Derivation of Maximal Common Subgraphs of Two Undirected Graphs". *Calcolo* **1972**, *9*, 341.

(28) Jansen, P. A. In "International Encyclopedia of Pharmacology and Therapeutics"; Cavalito, C. I., Ed.; Pergamon Press: New York, 1975; Vol. 1, Section 5.

(29) Piruzian, L. A.; Avidon, V. V.; Rozenblit, A. B.; Arolovich, V. S.; Golender, V. E.; Kozlova, S. P.; Mikhailovsky, E. M.; Gavrishchuk, E. G. "Statistical Analysis of the Information File on Biologically Active Compounds. I. Data Base on the Structure and Activity of Biologically Active Compounds". *Khim.-Farm. Zh.* **1977**, *11* (4), 35–40 (in Russian).

(30) Avidon, V. V.; Arolovich, V. S.; Kozlova, S. P.; Piruzian, L. A. "Statistical Study of the Information File on Biologically Active Compounds. II. Choice of Decision Rule for Biological Activity Prediction". *Khim.-Farm. Zh.* **1978**, *12* (5), 88–93 (in Russian).

(31) Avidon, V. V.; Arolovich, V. S., Kozlova, S. P.; Piruzian, L. A. "Statistical Study of the Information File on Biologically Active Compounds. III. Prediction of the Biological Activity Using Substructural Analysis Methods". *Khim.-Farm. Zh.* **1978**, *12* (6), 99–106 (in Russian).

(32) Gavrilova, V. V.; Golender, V. E.; Rozenblit, A. B.; Sukhova, N. M.; Lidaks, M.; Lukevics, E. "Statistical Analysis of the Information File on Biologically Active Compounds. IV. Prediction of the Biological Activity of Chemical Compounds Using an Algorithm Based on a Logical Structural Approach". *Khim.-Farm. Zh.* **1979**, *13* (2), 45–53 (in Russian).

(33) Gitlina, L. S.; Golender, V. E.; Rozenblit, A. B. "The Method for the Selection of Topological and Topographical Features of the Biological Activity of Chemical Compounds". *Khim.-Farm. Zh.* **1980**, *14* (7), 48–52 (in Russian).

(34) Gitlina, L. S.; Golender, V. E.; Drboglav, V. V.; Rozenblit, A B; Eihenberga, R. A., Avidon, V. V. "Methods for Representation and Processing of Structural Information for Structure–Activity Relationships Analysis". Preprint by the Institute of Organic Synthesis, Riga, USSR, 1981 (in Russian).

(35) A cover-to-cover English translation of "Khimiko-Farmatsevticheskii Zhurnal" is published by the Consultants Bureau, New York, under the title *Pharmaceutical Chemistry Journal*.