Theoret. Chim. Acta (Berl.) 58, 45–68 (1980)

THEORETICA CHIMICA ACTA
© by Springer-Verlag 1980

A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations

Charles L. Wilkins and Milan Randić [1]

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588, U.S.A.

The structure similarity and dissimilarity implied in many structure-property and structure-activity relationships has been examined from the graph theoretical point of view. The approach outlined is fundamentally different from generally used schemes in that, rather than seeking a new parametrization which will quantitatively fit observed data and trends, *similarities* among the skeletal forms and connectivities of the compounds of interest are studied quantitatively. The basis of the method is the assumption that skeletal forms of apparent similarity will yield similar enumerations for a number of graph theoretical invariants. In particular, all *paths* within molecular skeletons are enumerated and sequences of path numbers (i.e., the number of paths of different length) are compared. The degree of similarity between molecules is proportional to the distance between points in the corresponding "structure space" obtained by interpreting the entries in molecular path sequences as coordinates in n-dimensional space. As an *example* of the use of the concept of structural similarity, structure-activity data relating cerebral dopamine agonist properties for a series of N-substituted 2-aminotetralins are considered. The analysis suggests that the method may find wide application in the field of structure-activity correlations and structure-property studies. The data could be mass spectra, the "fingerprint" regions of infrared spectra, optical rotation and circular dichroism measurements, or any of many not fully-understood complex experimental findings suspected of having an inherent structural basis.

Key words: Structural similarity – Graph-path enumeration – Dopamine agonist properties – Molecular connectivity.

0040-5744/80/0058/0045/$04.80

1. Introduction

Similarity among structures is frequently invoked in discussions of molecular properties although it has been always used in a qualitative and unspecific manner. Even in such qualitative form, the concept has led to many important conclusions, particularly in structural chemistry. For example, the mass spectra of tricyclene (I) and cyclofenchene (II) (Chart 1) show remarkable similarity, reflecting the similarity of their molecular skeletons [2]. This may be contrasted with mass spectra of other related tricyclic terpenes (such as adamantane). The similarity of the infrared spectra of two derivatives of quinolizine (III), (IV) (Chart 1) provides another illustration [3]. Even in the absence of a detailed assignment of the bands

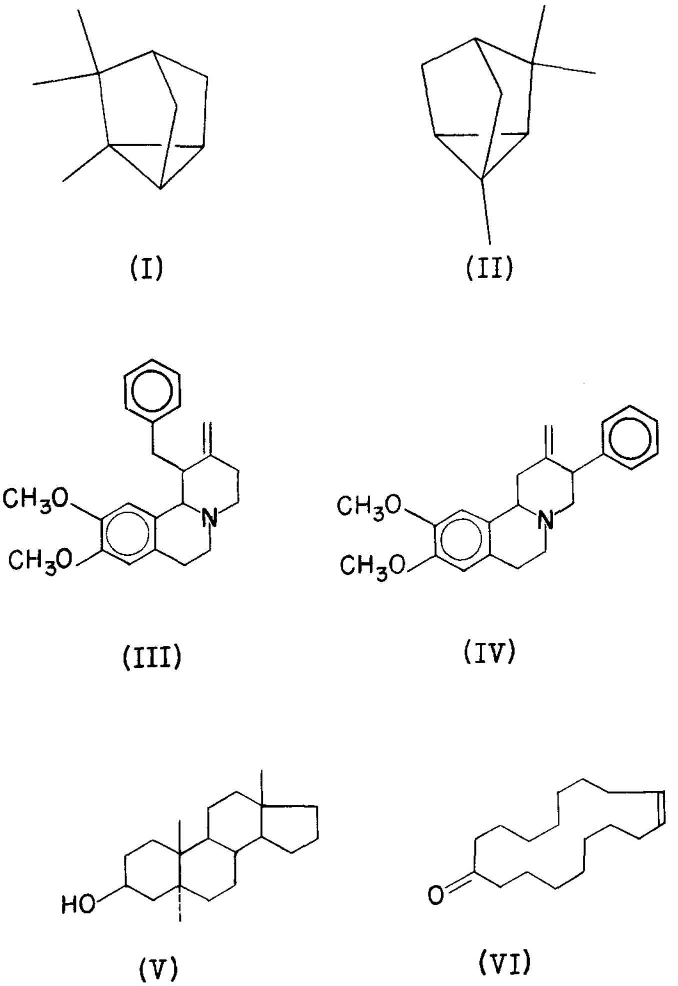


Chart 1. Skeletal forms of selected compounds showing great similarity in some of their properties: *(I)* amd *(II)* similarity in mass spectra; *(III)* and *(IV)* similarity in infrared spectra; *(V)* and *(VI)* similarity in odor

the apparent similarity strongly suggests common skeletal features. When comparing structures for similarity it may be that some parts make the dominant contribution. Consider the observation of Prelog and Ružička [4] of the remarkable formal resemblance between the structure of a sterol (V) which possesses a decidedly musk-like odor and the skeleton of the macrocyclic musk, civetone (VI) (Chart 1). Finally, the *lack* of similarity is sometimes instructive as well. The lack of similarity in the infrared spectra of ethane and diborane indicated that the molecules do not belong to the same symmetry point group, which supported the bridged hydrogen structure for diborane [5]. The range of problems where similarity plays an important role is very broad and the examples selected only serve to suggest the scope of applications in which benefit is expected if the concept of structural similarity is quantified. We are concerned in this paper with rigorous characterization of structural similarity among molecular skeletons and particularly address the application of a new quantitative structure similarity metric to structure-activity correlations. Since structure-activity correlations frequently imply some structure-property relationship, it is apparent that our approach is equally well-suited for the study of structure-property relationships. It is this possible wide applicability of the approach which prompts us to communicate the outline of the method at this initial stage of development. It may well be that subsequent applications to diverse problems will result in modifications of lesser generality, so it now seems proper to delineate the essential elements of our graph theoretical method as it may apply to quantitative structure-property and structure-activity correlations.

2. On Structure-Activity Correlations

Although the objective of reaching a detailed understanding of the interaction of drugs with drug-receptor sites remains the ultimate goal of structure-activity studies, the complexity of the problem has demanded the use of simplified models, simulations and analogies. Practical schemes for the study of structure-activity relationships continue to be useful and of interest. In Hansch's approach [6], empirical parameters are selected for prediction of $\log(1/C)$, where C is the molar concentration of the active substance causing 50% inhibition of the corresponding biological response. Free and Wilson [7], on the other hand, use the assumption of additivity of contributions of selected substructures on overall biological activity. Here even the same functional group, if located in different parts of a molecule, may be assigned different contributions. Both these methods are analogous to curve-fitting procedures employing many parameters. In contrast, correlations based on the concept of the connectivity index [8] use as a rule very few adjustable parameters, which at the same time have a simple structural interpretation. As demonstrated, [8, 9] impressive correlations with high accuracy can be achieved, primarily because a preselected bond weighing was determined so that ordering of structures parallels ordering of properties. Finally, the techniques of pattern recognition have been applied to classification of therapeutic agents [10]. Here, essentially one seeks groups of compounds of similar activity which are discriminated using a suitable set of weighting factors, empirically determined with the help of computer programs optimized previously with an adequate training set of compounds.

Implicit in all these methods is the assumption of additivity of structural effects, although, strictly speaking, preprocessing methods can introduce non-linearities. The methods differ in their mode of selecting the critical elements. That atoms, bonds, functional groups, and fragments are useful constituents in searching for an *additive* property is too well-known to merit argument. In connection with biological activity and pharmacological activity it is, however, also generally accepted that molecules of *similar* structural form may be expected to show *similar* biological or pharmacophoric patterns. It is self-evident that in structurally similar systems (the term *similar* is used here in its intuitive connotation) one can expect to find similar types and numbers of constituent fragments. The converse is *not* necessarily true. Compounds of vastly different form and shape can be built from the same building blocks. Hence, when *additivity* is used as the basis for quantitative structure-activity correlations instead of *similarity* we may have already fragmented the contributions too much, and thus permit inclusion of structures which need not be relevant for the particular study but happen to have same constituents. We hypothesize that if the search for regularities were based upon *quantitative similarity tests*, a number of unproductive compounds would have been eliminated from consideration. In this paper, we will examine the concept of similarity closely and, thus, try to supplement existing schemes with a quantitative similarity test for structures of interest. We will not produce additional parameters for an alternative additivity approach, but rather, will focus our attention on *quantification of structural differences among molecular skeletons*. For a pair of structures, we can estimate the degree of their similarity (or dissimilarity) and express that estimate as an index. When the degree of similarity is large, one can expect that many physico-chemical molecular properties will be similar. If the molecules are biologically active or therapeutically useful, they will show related activity patterns. The notion of similarity and its mathematical characterization is not so novel [11]. It has been found useful in biology. For instance, it has utility in relating protein amino acid sequence data of different species [12]. In chemistry, in a series of papers Dubois [13] has considered structural differences systematically.

Typically, in order to compare structures, graph theoretical approaches [14] represent a molecule by a sequence of integers, numbers usually obtained by enumeration of selected graph invariants. Comparison of *sequences* requires criteria to be followed in resolving which sequence dominates others, if they can be compared at all. The problem occasionally arises in the literature and has led to formulation of the rules for ordering sequences [15]. However, it appears that this particular problem was first considered at the beginning of this century by Muirhead [16], and has been further developed and generalized [17]. The problem of similarity between structures represents an additional application of the mathematical comparison of sequences. The characterization of a structure which we will use here is based on enumeration of *paths* (*vide infra*) which appears to agree with our intuitive perception of molecular similarity [18]. This approach was found useful in screening out an excess of computer-generated structures when fewer would have been desirable. Besides serving such an important editing role, our similarity measure has also revealed that, among numerous hypothetical structures, such as hypothetical monocyclic monoterpenes, *those actually found in nature show a relatively high degree of similarity among themselves* [18]. *This is an important finding* suggesting that biological generation is rather specific and very selective with respect to the plethora of mathematically allowed forms. Although the method does not distinguish the steric factors in such biosynthesis, having a method to establish the degree of similarity among the structures we are in a position to search for active substances more efficiently and more intelligently than without such a tool. Here, we illustrate the approach with a dozen aminotetralins, for the majority of which the stereotypical behavioral effects on experimental animals have been reported. The fundamental premise of the relationship between structure and activity, when structural similarity is tested *quantitatively* as outlined here, has been found to hold.

3. Outline of the Similarity Test

In order to give to the concept of similarity a quantitative meaning, we must not only define what *similarity* is (and how to measure it); but we must also define the structures in terms of *quantities* that can be subject to a quantitative estimate. In fact, the two tasks parallel each other, since *similarity* is not an absolute but a relative concept: one has to specify to what attribute the similarity is attached (e.g., similarity in shape, similarity in size, similarity of spectra, similarity in metabolic pathways, similarity of odor or taste, and so forth). Here we are concerned with similarity in the molecular connectivity (i.e., similarity of molecular graphs) [19]. Accordingly, we do not discriminate between the types of bonds and the kinds of atoms. Equally, we do not differentiate stereochemical features. By adopting these rather drastic simplifying properties of graphs, one might expect to limit the utility of the scheme for practical problems. Clearly the very factors we have ignored are of crucial importance to molecular activity. As we will see, the difficulties can be obviated by judicious selection of compounds of the same stereochemical configurations and similar molecular composition. In other words, other things being equal, we are mostly interested in examining the role of the molecular connectivity and to what extent different behavior of compounds can be traced to such topological differences between the compounds. The same constraints are often implicitly incorporated in other more conventional schemes.

Operating within these guidelines, we consider molecular graphs and enumerate all paths (i.e., self-avoiding walks) of different length. A self-avoiding walk or path is given by a sequence of consecutive edges in which no vertex appears more than once. The concept is illustrated in Table 1 for the structure of Fig. 1, which represents one of the aminotetralin derivatives considered in this paper. There are only two paths of length one, since only two bonds meet at vertex one. There are three paths of length two, since at the adjacent vertex 10, it is possible to branch either to vertex 5 or vertex 9. The number of paths of length three is still larger

Table 1. All paths starting with atom 1 grouped according to the length of the path. Total number of paths in each group is shown at the right and constitutes the corresponding atom code

Path length	Paths	Number of paths
1	1-2 1-10	2
2	1-2-3 1-10-5 1-10-9	3
3	1-2-3-13 1-2-3-4 1-10-5-4 1-10-5-6 1-10-9-8 1-10-9-11	6
4	1-2-3-13-14 1-2-3-13-15 1-2-3-4-5 1-10-5-4-3 1-10-5-6-7 1-10-9-8-7 1-10-9-8-12	7
5	1-2-3-4-5-6 1-2-3-4-5-10 1-10-5-4-3-2 1-10-5-4-3-13 1-10-5-6-7-8 1-10-9-8-7-6	6
6	1-2-3-4-5-10-9 1-2-3-4-5-6-7 1-10-5-4-3-13-14 1-10-5-4-3-13-15 1-10-5-6-7-8-12 1-10-5-6-7-8-9 1-10-9-8-7-6-5	7
7	1-2-3-4-5-10-9-11 1-2-3-4-5-10-9-8 1-10-5-6-7-8-9-11 1-10-9-8-7-6-5-4 1-2-3-4-5-6-7-8	5
8	1-2-3-4-5-10-9-8-12 1-2-3-4-5-10-9-8-7 1-2-3-4-5-6-7-8-9 1-2-3-4-5-6-7-8-12 1-10-9-8-7-6-5-4-3	5
9	1-2-3-4-5-10-9-8-7-6 1-2-3-4-5-6-7-8-9-10 1-2-3-4-5-6-7-8-9-11 1-10-9-8-7-6-5-4-3-2 1-10-9-8-7-6-5-4-3-14	5
10	1-10-9-8-7-6-5-4-3-13-14 1-10-9-8-7-6-5-4-3-13-15	2

since we reach additional branching sites (3, 5, 9). We depict in Fig. 2 all paths of length four for atom 1. In polycyclic structures the number of paths proliferates rapidly with the increase in the number of rings, and it soon becomes impractical, except for relatively simple systems, to visually enumerate them. A program for enumerating (and if desired, listing) of paths of different length is available [20]. A typical output appears in Table 2. Here, paths of different length for all the atoms of the aminotetralin shown in Fig. 1 are listed. Notice that in the molecular graphs contained in Figures 1–6, the nitrogen atoms (always at the 2 position), the hydroxyl groups (all mono- and disubstituted aromatic rings have *only* hydroxyl groups), and aromatic rings (the rings at left) are not explicitly identified. Enumeration of paths of all possible lengths results in attribution to each atom a sequence of integers. The sequences, when suitably truncated, are of interest in studies of local atomic properties (e.g., carbon-13 chemical shifts in NMR [21]. In the present context we would prefer some reduced information which would

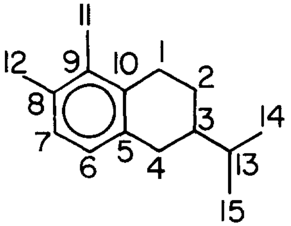


Fig. 1. Molecular graph of dimethyl substituted 2-amino-5,6-dihydroxytetralin, one of the compounds investigated. The ring at left is aromatic. (Numbering is arbitrary.)

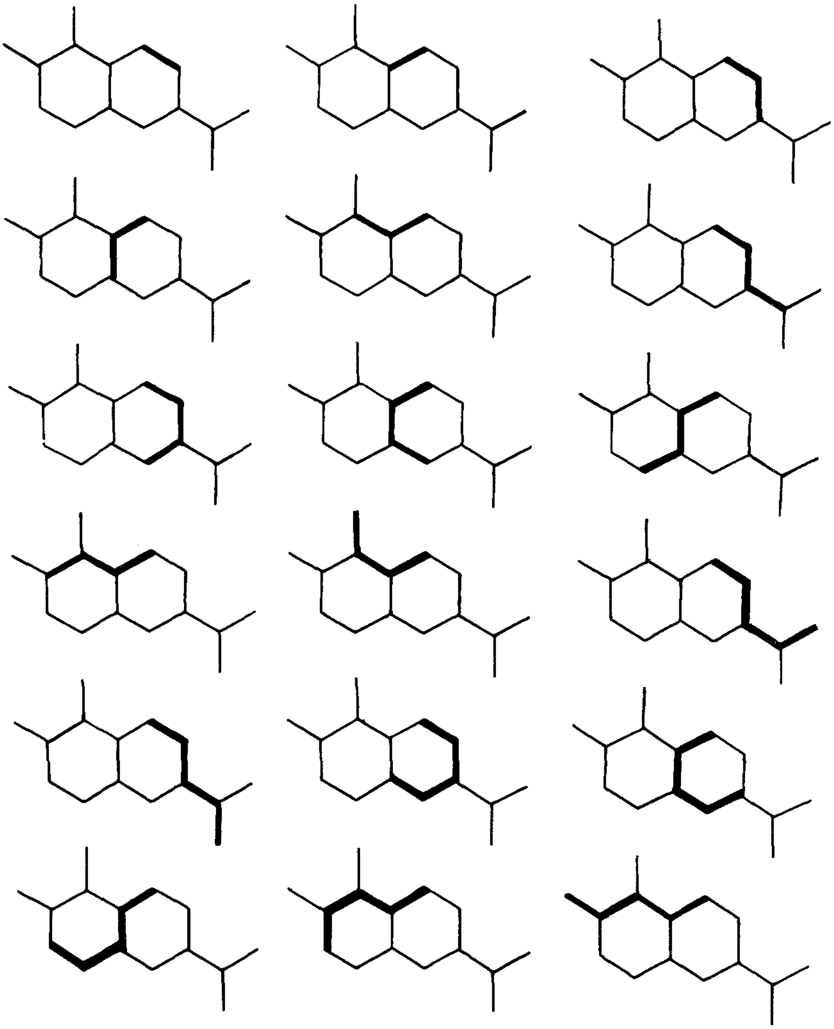


Fig. 2. Illustrations of all paths of length four originating with atom 1. The substituents on the left ring are hydroxyl, the right ring substituent is dimethylamino

pertain to a *molecule* as a whole, not to the collection of individual atoms. The most natural contraction is obtained by summing atomic sequences, term by term, to form a single molecular sequence. The last row in Table 2 gives the result of the summation, which has been divided by two, because each path appears twice in the sum, once for each end atom. The entries in the derived molecular sequence give the number of paths of different length in the molecule as a whole. Thus, there are 16 paths of length one (i.e., 16 bonds); 23 paths of length two (i.e., 23 pairs of adjacent bonds); 30 paths of length three, 36 paths of length four, and so on. Although we have arrived at a *molecular* characterization by *adding* atomic contributions observe that here the atomic contributions are not *local* in character

Table 2. Paths for all the fifteen atoms in the aminotetralin derivative (Fig. 1) listed in order of increasing path length. Zero path length corresponds to the count of atoms. The molecular sequence is derived by summing all the contributions for atoms and dividing the result by 2, since each path has been encountered twice (once for each end atom), except for the first column which is a simple sum

Path length:	0	1	2	3	4	5	6	7	8	9	10	11
Atom												
1	1	2	3	6	7	6	7	5	5	5	2	
2	1	2	3	5	5	7	6	9	5	2	1	2
3	1	3	4	3	5	8	7	5	4	3		
4	1	2	4	6	5	7	7	6	3	4	2	2
5	1	3	4	6	8	6	4	3	3	3	2	
6	1	2	3	5	7	8	5	5	6	8		
7	1	2	3	4	5	8	7	7	9	3	1	
8	1	3	3	3	5	6	8	10	4	4	1	
9	1	3	4	4	4	7	9	4	5	4	3	
10	1	3	5	5	6	9	3	2	3	5	1	
11	1	1	2	4	4	4	7	9	4	5	4	3
12	1	1	2	3	3	5	6	8	10	4	4	1
13	1	3	2	2	3	5	8	7	5	4	3	
14	1	1	2	2	2	3	5	8	7	5	4	3
15	1	1	2	2	2	3	5	8	7	5	4	3
Molecule												
	15	16	23	30	36	46	47	48	40	32	16	7

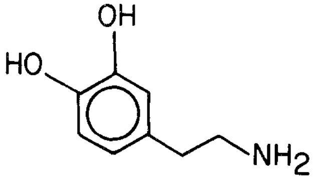
but contain information on *all* distant neighbors. This differentiates our approach and a few similar schemes [22] from the customary additivities in which terms carry local character.

The use of path numbers was anticipated long ago by Platt [23] in connection with a study of isomeric variations in the physio-chemical properties of alkanes. Altenburg has expressed the mean squared molecular radius (for alkane isomers) in terms of polynomials in which path numbers are the coefficients. One can also use path numbers for ordering of structures. Illustrations are provided in discussions of regularities in thermodynamic properties of octanes and higher alkanes [24]. In the present application, we are interested in the similarity among structures. The *quantitative* aspect of our approach is the *computation* of the degree of similarity (or dissimilarity) among various structures. When referring to accompanying molecular properties, the approach is *qualitative* since we normally infer from structures that properties show some parallelism. When a large number of related molecules are investigated, however, one may attempt some quantitative discussion of the relative magnitudes of the property of interest by interpolating or extrapolating from few well selected samples. What is rigorous in our approach is the ranking or partial ordering of structures designed to parallel the ordering of the activity of the compounds.

Since frequently molecules of different size must be compared, some normalization scheme appears desirable so that the difference in size alone does not dominate the ordering. There are several plausible normalization procedures: (1) one can divide enumerated paths by the number of atoms in the structure; (2) one can divide path numbers by the number of paths in a structure; or (3) one can normalize individual atomic path sequences (by dividing them by the number of atom paths) prior to summation into molecular sequences. Each procedure emphasizes a somewhat different structural aspect. With the first normalization, which corresponds to use of average atom codes, the effect of a given difference is relatively more pronounced in small molecules than in larger ones. If the second normalization is used, the differences between structures of the same size are de-emphasized. The third alternative has the effect of causing exocyclic substituents to make greater relative contributions than internal (cyclic) atoms to the similarity measure. One can also conceive of other scaling schemes which would range from no normalization at all, (i.e., taking the sequences of path numbers directly regardless of the size of a molecule) to the most general scheme, which would weight paths according to a preconceived plan. However, it would be premature to explore such considerations before establishing how the approach works in its simplest form. For this reason, unnormalized path codes and the simplest normalization based on the average atom path numbers were examined first [25, 26].

4. An Illustration

As an illustration, we consider a selection of variously N-substituted 2-aminotetralins having OH groups at the 5 and 6 or 6 and 7 positions. These compounds are similar to dopamine, which suggested their testing for dopaminergic agonistic potency [27]. The molecular graphs for the compounds considered, which induced behavioral effects in experimental mice (sniffing, compulsive gnawing, and hyperactivity) are shown in Fig. 3. In this case, activity could be graded on a scale 1–4, the compounds varying in the dose (mg/kg) necessary to induce stereotypic activity [2, 7]. The related compounds of Fig. 4 differ in their substitution pattern and are apparently similar. Yet, the first group (Fig. 3) is active while the second group (Fig. 4) is inactive [27]. The problem we wish to consider is whether path enumeration can assist in discriminating the two groups and whether it can quantify otherwise minor differences among the compounds shown. In Table 3, we list path enumerations for the structures of Fig. 3 and Fig. 4 as obtained using our computer program for enumeration of paths [20]. The active molecules have been ordered alphabetically according to their relative dopaminergic activity: A is the most potent, B, C, and D are relatively potent and the others somewhat less



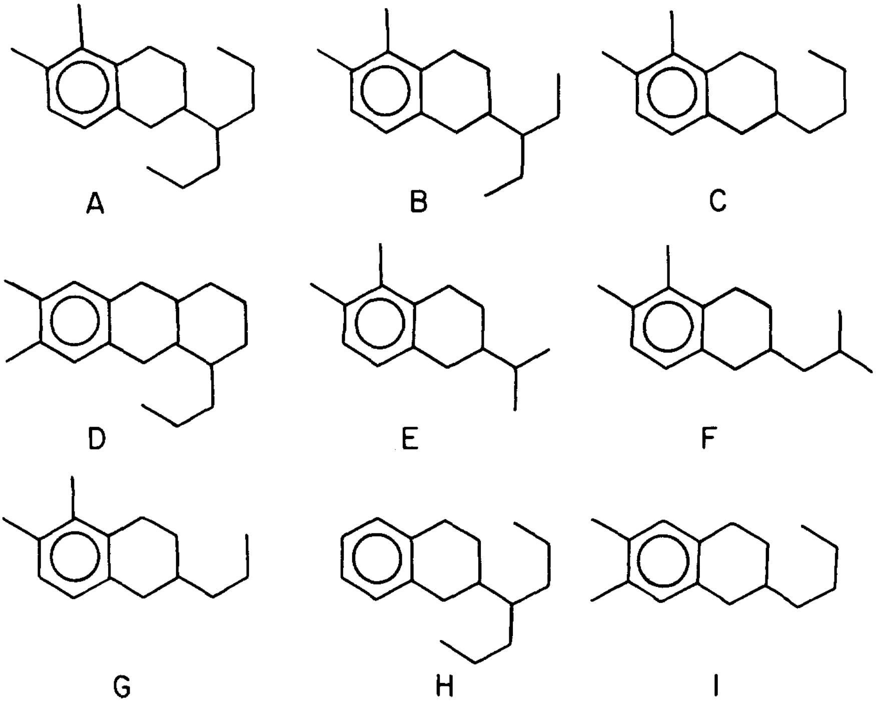


Fig. 3. Molecular graphs for aminotetralin derivatives showing dopaminergic potency by inducing typified behavioral effects in experimental mice. All aromatic ring substituents (at left) are hydroxyl; the alicyclic ring substituents (at right) are mono- or dialkylamino groups

potent. Compounds J–R lack the activity described above. As is seen from Table 3 the differences among various structures become more pronounced with paths of longer length. Therefore, it is important in the analysis of the similarity to retain *all* paths and not truncate the sequence. In Fig. 5 we represent the path sequences for selected molecules in a pictorial form from which one can visualize the different character of the individual sequences. Here the path numbers are shown as "intensities", the abscissa being the length of the paths. The resulting diagrams, which may be called "path spectra", translate the similarity among the sequences into the similarity of the line distributions. Such a representation contains the same information as path lists, but sometimes the differences among the structures are more readily apparent in the pictorial representations than in the tabular form.

Using the data of Table 3 on path counts we can define as the measure of similarity for a pair of structures the associated *distance* between the structures in a "structure-space" when path numbers are interpreted as coordinates of a structure in multidimensional vector space. Let $p_i(\mathrm{A})$ and $p_i(\mathrm{B})$ represent path numbers of length i for structure A and structure B, respectively. The distance is given then by:

$$\mathrm{D(A, B)} = \sum^{n} [(p_i(\mathrm{A}) - p_i(\mathrm{B})^2]^{1/2}$$

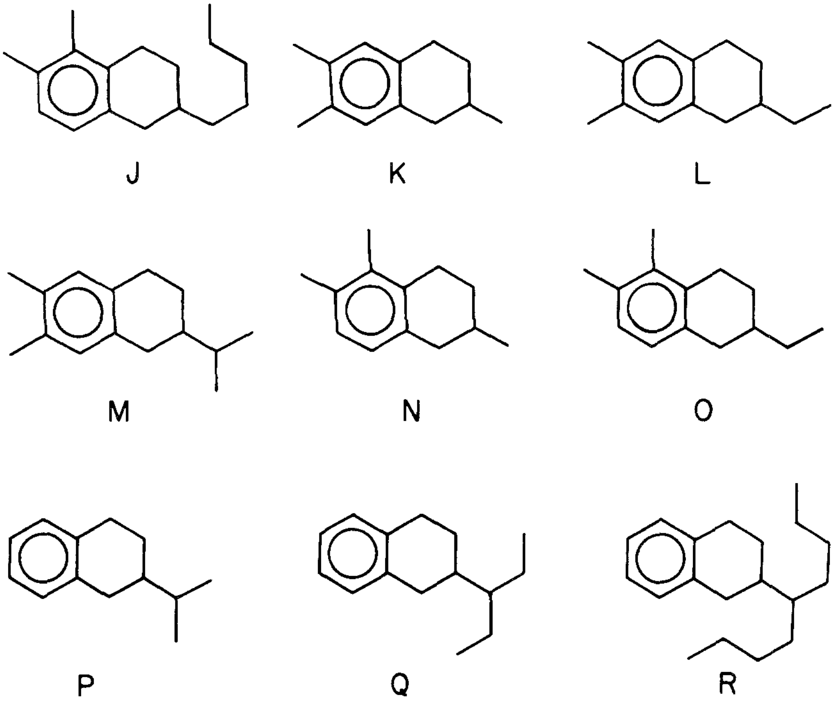


Fig. 4. Molecular graphs of additional 2-aminotetralin derivatives found inactive in experimental animals. Substituents are as explained in Fig. 3

Table 3. Molecular path sequences for the selected derivatives of 2-N-aminotetralins shown in Fig. 3

Path length:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Molecule															
A	19	20	27	36	45	56	58	64	66	62	40	25	14	6	
B	17	18	25	34	41	50	53	58	56	46	26	15	6		
C	16	17	23	30	37	47	47	48	46	42	24	13	7	3	
D	19	20	27	35	44	57	58	62	67	64	42	27	12	4	
E	15	16	23	30	36	46	47	48	40	32	16	7			
F	16	17	24	30	38	47	48	50	49	41	22	12	6		
G	15	16	22	29	36	45	45	45	41	34	17	8	3		
H	17	18	23	30	38	47	45	48	48	43	24	18	10	4	
I	16	17	23	29	36	48	47	47	48	43	24	14	6	2	
J	17	18	24	31	38	48	49	50	49	47	32	20	12	7	3
K	13	14	20	25	31	41	37	32	29	22	6	1			
L	14	15	21	27	33	44	42	39	36	29	10	3			
M	15	16	23	29	35	47	47	46	43	36	14	5			
N	13	14	20	26	32	40	37	33	26	22	8	1			
O	14	15	21	28	34	43	42	40	33	27	12	4			
P	13	14	19	24	29	37	34	32	26	8	4				
Q	15	16	21	28	34	41	40	42	38	31	16	10	4		
R	19	20	25	33	40	51	51	54	55	53	36	26	18	10	4

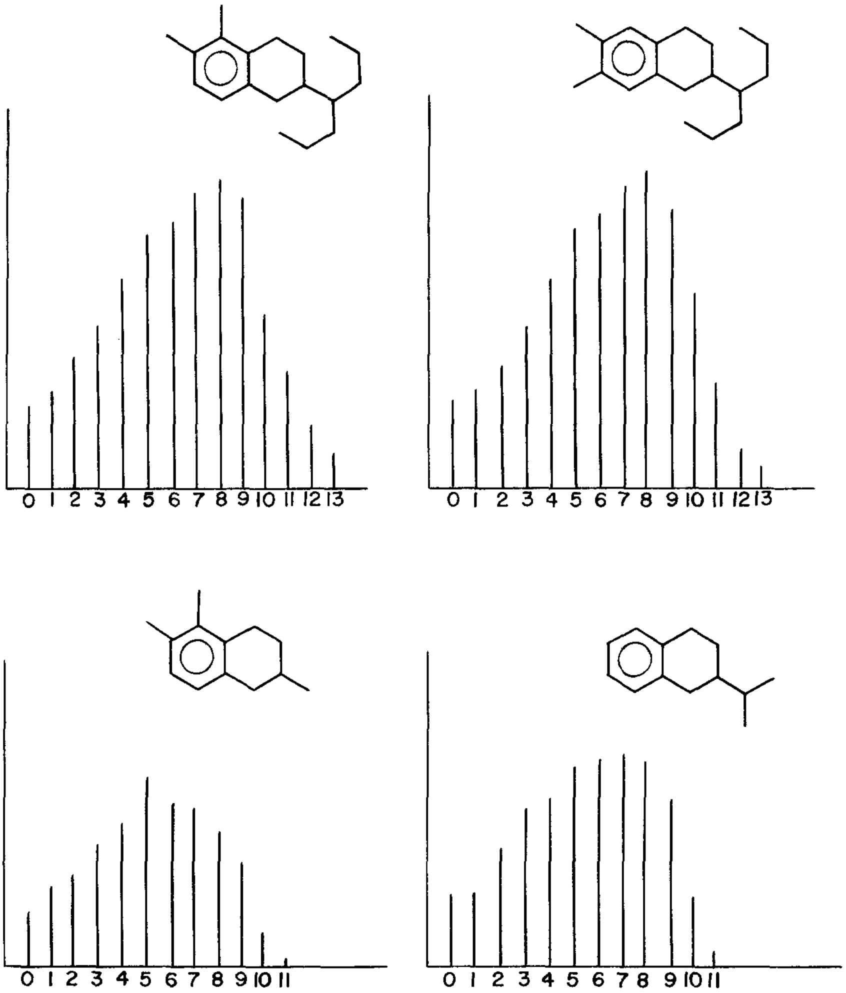


Fig. 5. Pictorial representation of path numbers for selected aminotetralin derivatives. Substituents are as explained in Fig. 3

where the sum extends to the larger of the n's, n being the number of entries in the sequences. In Table 4 are the distances for all the pairs of structures of Figs. 3 and 4. The entries are part of an 18 × 18 symmetrical matrix. The upper part of Table 4 corresponds to active compounds, while the lower part of the same table shows distances between inactive and active compounds. The similarities between various inactive compounds are of no interest here and were disregarded. Close examination of Table 4 reveals several interesting details. We observe that pairs of compounds which are *very* similar, such as the pairs (A, D), (C, F), (C, H), (C, I) show parallel differences from third structures. This parallelism confirms that the particular structural encoding based on path enumerations satisfies the requirements of metrics for "structure space". Strictly speaking the requirements for a

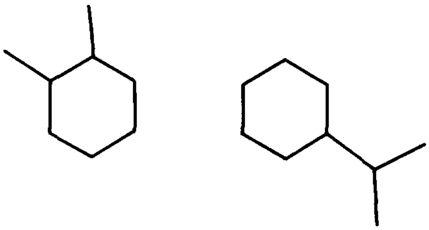
Table 4. Part of the dissimilarity matrix. Each entry represents the distance of the points defined by sequences of Table 3. The upper triangular part gives the dissimilarity coefficients among active compounds, while the lower part corresponds to distances between agonists (compounds A–I) and compounds reported to be inactive (compounds J–R of Fig. 4.)

Molecule									
A	*								
B	29.3	*							
C	43.0	17.5	*						
D	4.9	31.8	44.7	*					
E	57.9	29.3	17.3	68.9	*				
F	42.8	15.0	5.7	44.3	16.5	*			
G	56.8	28.9	14.5	58.5	5.7	14.6	*		
H	40.3	17.9	6.9	42.0	21.5	9.6	19.3	*	
I	42.2	17.3	3.5	43.5	15.0	5.7	15.7	7.0	*
	A	B	C	D	E	F	H	I	J
J	32.4	17.5	14.5	34.0	31.0	17.3	22.1	11.4	14.5
K	84.9	56.5	41.2	70.2	26.2	38.8	27.7	43.8	41.1
L	70.3	42.3	27.8	71.9	14.0	28.0	14.0	32.1	29.2
M	56.9	27.0	16.7	58.6	6.3	15.2	6.8	22.0	17.3
N	84.1	56.7	41.2	85.8	28.4	42.3	27.9	44.9	30.8
O	71.0	43.1	28.4	72.4	14.5	29.0	14.5	32.7	29.7
P	85.5	58.5	42.6	87.0	31.1	43.9	29.9	45.6	43.6
Q	63.0	36.0	20.4	64.6	12.3	21.4	8.9	23.5	21.6
R	21.9	23.7	28.4	23.1	41.0	30.1	40.2	24.2	28.1

metric imply that the distance D is positive, does not depend on the direction of measurement, is definite, and satisfies the triangular rule, (i.e., the direct route gives the shortest distance) [28]. The definiteness, D(X, Y) = 0 implying X = Y, while being a condition for metrics is sometimes not required in considerations of dissimilarity coefficients [29]. The possibility that D(X, Y) = 0, while $X \neq Y$, amounts to the situation that the "structure space" represents some projection of more complete space. As long as the overlooked part of the structure space carries negligible or not very relevant content such distortions from strict metric behavior will hardly matter. If we were disregarding some important part of the structural characterization two structures found similar could in fact have additional significant differences, but structures already found different could only increase their lack of resemblance. Therefore we can draw inferences about the structural similarity between the compounds and expected properties and use the similarity distances even indirectly, by observing how two compounds compare with a number of other structures of interest. On this basis, if one compares the various rows and columns of Table 4, (using also the information relating to inactive compounds), it can be established that there are possibly three categories of active compounds as follows:

I A, D
II B, C, F, H,
III E, G

with the last two groups showing minor differences. It appears that the structure D derives its appreciable activity from its high similarity with the most active structure A. The structures F, H, I, and possibly E and G seem to be active because of significant similarity with structures B and C, which themselves are highly potent, yet appear to have significant structural differences from structure A. The limited similarity among A and B or A and C suggests a possibility that more than one structural feature may be responsible for the manifestation of the particular biological activity. Moreover, one can attempt to identify the fragments presumably responsible for the activity. With E being similar to G, an active compound, and M, an inactive compound, we can deduce that of the two ring fragments that constitute E:



the former is more dominant for the activity. Therefore if modifications based on the compound E are contemplated, it is the second fragment that should be the primary candidate for alteration. Such information would be of interest in drug design, even if it does not directly suggest what alterations should be considered. We already benefit by eliminating numerous other possibilities that would involve the dominant fragment and which are likely to be unproductive by simply spoiling the already desirable features of the dominant fragment.

The prevailing role of size as a singularily important fact in this comparison of structures cannot be denied. If we order the compounds A–I according to the magnitude of their similarity with A, we observe that similarity parallels the size, indicated by the number of "heavy" atoms [30].

	D	B	H	I	F	C	G	E
similarity distance to structure A (approx.)	5	30	40	40	40	40	60	60
size (i.e., the number of "heavy" atoms)	19	17	17	16	16	16	15	15

While it is understandable that the size of a structure should be a very important factor in structure-activity correlations, it would also be of interest to suppress this particular feature in some comparisons so that *other* factors of interest become visible. For example, the most active available compound, assumed as the standard, may already contain minor details *not* relevant for the particular study. The presence of such an irrelevant fragment increases the size of the structure and consequently favors structures of similar *size* at the expense of structures which may have a closer resemblance to the standard should the irrelevant details be ignored. By normalization one can de-emphasize the role of the size. We will consider average path numbers, derived by dividing the numbers obtained by enumeration by the number of atoms in the corresponding graph. For the compounds A and B, we then obtain respectively:

A	1.05	1.42	1.89	2.37	2.95	3.05	3.37	3.47	3.26	2.11	1.32	0.74	0.32
B	1.06	1.47	2.00	2.41	2.94	3.12	3.41	3.29	2.71	1.53	0.88	0.35	

The distance between the two structures is given by

$$D(A, B) = [(0.02)^2 + (0.06)^2 + (0.05)^2 + \cdots + (0.32)^2]^{1/2} = 1.06.$$

The upper part of Table 5 is half of a symmetric 9 × 9 table in which the degree of similarity among the dopaminergic agonists is indicated. The lower part of the same table reveals how similar (or dissimilar) to the potent substances other inactive 2-aminotetralin derivatives are. A small value, such as 0.28 for the pair (A, D) or 0.22 for the pair (C, I) indicates the compounds of great similarity. An intermediate value, such as 1.96 and 1.91 for the pairs (A, E) and (A, G) respectively, points to compounds of some, but limited similarity. Such compounds will generally show disparate properties. A large value (such as 3.13 for the (A, N) pair), indicates considerable dissimilarity.

A lack of similarity among some of the most potent dopaminergic agonists of Table 5 does not necessarily signal inadequacy of our approach for the following reasons: first, the compound A may not be the ideal prototype of the hypothetical most active compound. Rather it may merely be the most potent compound

Table 5. Part of the dissimilarity matrix for compounds of Figs. 3 and 4 based on normalized molecular path numbers

Molecule									
A	*								
B	1.06	*							
C	1.32	0.68	*						
D	0.28	1.18	1.41	*					
E	1.96	1.14	0.97	2.07	*				
F	1.35	0.48	0.35	1.47	0.85	*			
G	1.91	1.00	0.59	2.03	0.38	0.60	*		
H	1.47	1.02	0.53	1.59	1.28	1.47	0.96	*	
I	1.23	0.65	0.22	1.30	1.05	0.36	0.77	0.56	*
	A	B	C	D	E	F	G	H	I
J	0.98	1.04	0.70	1.09	1.60	0.96	1.33	0.67	0.72
K	3.13	2.26	1.87	3.24	1.24	1.84	1.27	1.93	1.94
L	2.53	1.62	1.31	2.65	0.64	1.23	0.67	1.47	1.36
M	2.43	1.11	0.97	2.17	0.42	0.75	0.41	1.30	0.81
N	3.13	2.27	1.86	3.25	1.23	1.86	1.28	1.93	1.96
O	2.57	1.60	1.32	2.69	0.81	1.28	0.68	1.49	1.42
P	3.05	2.25	1.78	3.16	1.31	1.82	1.28	1.74	1.86
Q	2.16	1.39	0.95	2.27	0.83	0.96	0.60	0.86	1.01
R	1.15	1.41	1.07	1.25	1.97	1.35	1.69	0.83	1.06

among the present set. There may well be several other compounds having fair similarity with, say A and E (which is found the least similar to A), and displaying even better dopaminergic activity. In the next section, we will discuss possible structures for such compounds. Alternatively, it may be that two distinct structural features, one dominant in A and the other in E, are equally (or approximately equally) efficient in stimulating the biological effect. These are clearly open questions and further experimental evidence as well as some further theoretical refinements are required to pursue them. However, overall, one can observe that all the active compounds show a considerable degree of similarity, the distance of dissimilarity being less than or very close to 1.00 for most pairs. If we simply compare all the compounds with A (the first row in Table 5) we find a smaller average (1.32) for the active compounds than for inactive ones (average of 2.35). On this ground alone the compounds K and Q could be discarded as uninteresting, others being marginal, with only J and R displaying appreciable similarity with the standard. A similar conclusion follows from a comparison of inactive compounds with the standard B, the second most potent dopaminergic agonist among the set considered. The relative magnitudes for the average dissimilarity distances for the active (now 0.90) and inactive (now 1.66) compounds have not changed much, although the absolute numbers have decreased considerably. At this point, it appears that compound R no longer qualifies as attractive,while J still could be erroneously identified as potential agonist. The possible mistaken prediction that J could be a potent agonist is not alarming at all if one considers the overall objective of the type of similarity classification proposed here. *The end goal is an empirical tool for preliminary screening, prediction, and ranking of possible activity.* Such a method will serve well if it is found satisfactory at some high fidelity level. Observe also that some very definitive conclusions are possible, despite occasional ambiguity. Thus, one could eliminate the compounds K, N, and P immediately on the basis of their lack of striking similarity with *any* of the active compounds.

5. Ranking of Structures

We should bear in mind that we do not know the precise structural prerequisites for the specified activity and that more than a single structural feature may be of interest here. In addition, the effects of metabolism may further adversely affect compounds making the graph theoretical approach, which can probably adequately predict the partitioning of drugs (since it relates to surface area, a physical property), uncertain. One cannot deny such influences, but within a collection of relatively homogeneous populations, many such controlling factors may parallel the dominant features of the structure which are reflected in our graph theoretical approach. The question is whether the correlation with biological activity as implied in Table 5 may be coincidental. To answer the question we will consider a hierarchical ordering scheme, which we believe, demonstrates that the similarity basis developed here has some substance and the results are not fortuitous. The conclusions based on a comparison with a single standard may be deficient, being biased towards *all* structural components of the standard indiscriminantly. It is likely, in relatively large molecules, that some molecular features are more important for a particular property than others. A better discrimination of the most relevant structural features should result when considering comparisons with *several* standards simultaneously. In this way, the important structural details present in all selected standard compounds, will be weighed effectively at the expense of less pertinent fragments appearing in only a single structure. Such comparisons will generally result in *partial ordering* of structures or hierarchical ranking of the structures from which one can generally with greater reliance *predict* which among many candidate structures are more likely to show similar physico-chemical properties.

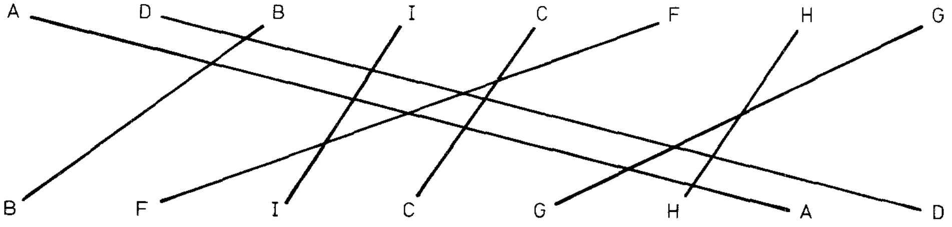
To clarify the ordering scheme, consider the active agonists of Table 5. Let us now sequence structures as dictated by A and B, (i.e., we order all the compounds by increasing similarity with structure A and B respectively):

A, D, B, I, C, F, H, G, E
B, F, I, C, G, H, A, D, E.

Now we compare the two orderings and seek all those subsets of the above sequences which preserve *both* orderings. They can be found by listing for each structure all the structures which follow it in *both* sequences. We immediately obtain:

A, D, E
B, I, C, F, H, G, E
C, H, G, E
D, E
E
F, H, G, E
G, E
H, E
I, C, H, G, E.

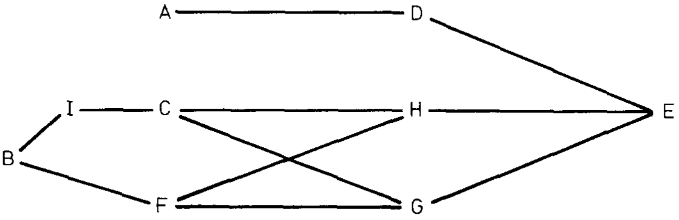
The above sequences contain redundant information and should be pruned by eliminating those sequences which are fully contained in larger ones. There is a simple way to extract the required information [31]. Write the initial sequences (starting with the structures A and B respectively) one above the other and connect the common letters. Each crossing of lines indicates the structures which cannot be simultaneously ordered with respect to A and B, while any sequence not



involving crossing of lines is an acceptable solution. Thus we immediately arrive at the following partial orderings which hold true for both A and B:

A, D, E
B, F, G, E
B, F, H, E
B, I, C, G, E
B, I, C, H, E.

One can summarize this information with a simple diagram placing structures that dominate others at the left:



The above graph depicts the hierarchical relationship among the structures defined by the similarity with respect to both A and B. From the diagram, we see that there may be two families of compounds with dopaminergic activity, A and D which form one branch of the diagram, and the remaining structures which appear in other branches. One can continue such an analysis by also including comparisons with structure C, which will further restrict compatibility among the structures, but the relatively small sample of the compounds considered here does not justify such additional efforts.

One should bear in mind that small differences in the dissimilarity matrix may not be significant, so when applying the above ranking of structures to a larger collection of compounds, structures may be grouped accordingly. Equally, one should be aware that this is not the only possible approach to characterization of structural similarity. However, it appears that the scheme outlined parallels our intuitive notions of structural similarity. This premise can be supported by considering a comparison of the structures, based on a count of how many bonds would superimpose each other when one molecular graph is placed above the other. In Table 6 we have summarized the results of such superpositions. Columns indicate the similarity distances of Table 5 grouped for the same number of non-overlapping bonds. Except for a few extreme values, we see that the similarity index of 0.3 characterizes structures with the same number of bonds, the values around 0.7 belong to structures with a single bond difference, values of 1–1.5 cover cases where the structures differ by two or three bonds and values over 2 belong to structures that differ by four bonds [32]. The major conclusion supported by the data of Table 6 is the substantial parallelism between a simple intuitive and coarse measure of similarity (based on the count of bonds in which structures differ) and the quantitative concept of similarity derived from the enumeration of paths of different length in a molecule.

Table 6. Illustration of the parallelism between an intuitive measure of the similarity among structures (based on the difference in the number of bonds when structures are superimposed) and a rigorous approach based on the corresponding coefficients of the dissimilarity matrix

	Bond Difference in Superposition of Structures				
	0	1	2	3	4
Singularity	0.217	0.478	0.975	1.266	1.905
of the	0.278	0.526	1.000	1.303	1.956
superimposed	0.354	0.564	1.064	1.319	2.028
structures	0.364	0.589	1.138	1.345	2.065
	0.377	0.598	1.180	1.406	
	1.019	0.653	1.279	1.469	
		0.680	1.469		
		0.700	1.548		
		0.769			
		0.845			
		0.973			
		1.052			

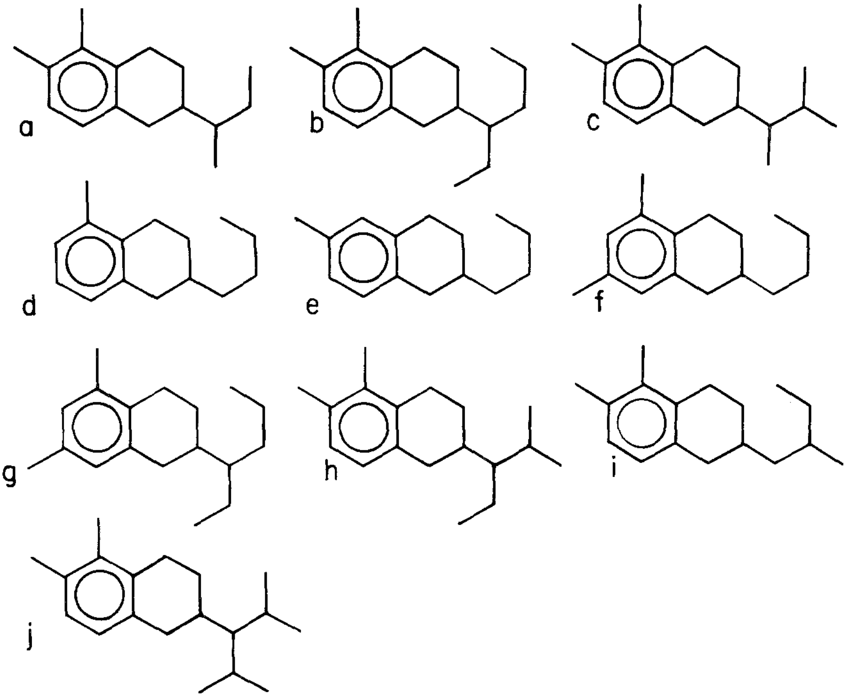


Fig. 6. Molecular graphs of hypothetical 2-aminotetralin derivatives screened for their potential dopaminergic activity. Substituents are as explained in Fig. 3

6. The Search for a Potential Novel Active Substance

A search for novel potential dopaminergic agonists may be based, as will be outlined, on comparison of hypothetical structures with already known agonists. We should, for that purpose, select the most potent compounds, such as A–C, because inclusion of less potent compounds is likely only to give prominence to structural factors which are less relevant to the sought-for activity. From Table 5, we see that *all* the active compounds can be detected as being similar to the three most active compounds, while the part of Table 5 corresponding to inactive substances is remarkably devoid of incidences of large similarity with A–C. On this premise, we examined a selection of hypothetical structures a–g, shown in Fig. 6 in order to predict their dopaminergic potency. In Table 7, we give path numbers and in Table 8 the dissimilarity coefficients for the new compounds. As we see, the entries in Table 8 are of the magnitude found for active compounds in Table 5. Hence we anticipate the compounds of Fig. 6 to be potential dopaminergic agonist. It is somewhat difficult to predict the relative strengths for the compounds since, as already stated, the structural prerequisites for the specific dopaminergic activity are not known. On the grounds that structures d and e show similar departures in corresponding distances from A and B, as the compound J, known to be inactive, we may expect the former may also have low activity, if any. In order to derive some insight into the relative activities (or potential activities)

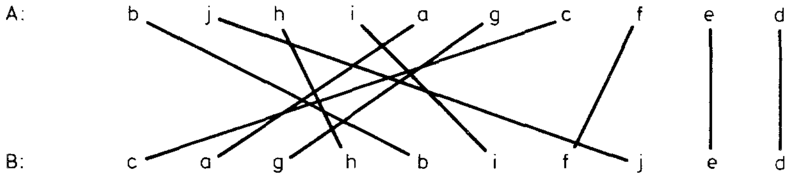
Table 7. Molecular path sequences for a collection of derivatives of 2-N-aminotetralins selected for prediction of their dopaminergic activity shown in Fig. 6

Path length:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Molecule															
a	17	18	25	33	40	50	52	56	53	47	28	16	7	3	
b	18	19	26	35	43	53	55	61	61	54	33	20	20	3	
c	17	18	26	34	40	50	53	58	56	46	26	15	6		
d	15	16	21	27	34	42	41	40	39	35	18	11	6	3	
e	15	16	21	26	33	43	40	41	40	36	19	10	6	2	
f	16	17	23	29	38	47	49	45	47	42	24	14	6	3	
g	17	18	25	33	42	50	55	56	55	45	31	12	6		
h	18	19	27	36	44	52	56	63	64	53	31	19	9		
i	17	18	25	32	39	49	50	53	54	49	29	17	10	3	
j	19	20	29	38	48	54	59	68	72	60	36	23	12		

Table 8. Part of the dissimilarity matrix between the most active dopaminergic agonists A, B, C, and the collection of aminotetralins tested for their potential biological activity

Molecule										
	a	b	c	d	e	f	g	h	i	j
A	0.97	0.50	1.21	1.88	1.62	1.31	1.09	0.72	0.89	0.62
B	0.36	0.57	0.08	1.26	1.19	0.75	0.40	0.48	0.57	0.91
C	0.47	0.89	0.67	0.66	0.60	0.26	0.70	1.00	0.48	1.40

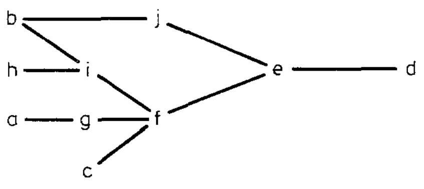
we ordered the data on the hypothetical compounds of Table 8 with respect to the most potent agents A and B, and again connected the common letters (structures):



Following the procedure already outlined for structures A–I, we obtain in the case of structures a–j five partial orders:

a, g, f, e, d
b, i, f, e, d
b, j, e, d
c, f, e, d
h, i, f, e, d.

These can be represented pictorially by the following graph:



Therefore, an intelligent selection as the most probable candidates for dopaminergic agonist activity is to be found among a, b, c, and h. If for example, a is investigated and *found* to be active (since as we have mentioned the approach represents a necessary but not a sufficient condition for predicting activity) this finding would suggest g as another viable candidate for testing. However, if h is similarly found potent, the next in line, the compound i is not necessarily of immediate interest, since we expect b to dominate whatever activity i may have. One must interpret such arguments with due caution, since similarity is based on the skeletal forms of active compounds rather than on some definitive requirements of the active site, which are unknown. Nevertheless, we see how with simple path enumerations and subsequent comparison, we can, surprising as it may appear, arrive at plausible conclusions on the potency of new compounds, and suggest a ranking of structures for planning testing, *even though the detailed mode of action and desired molecular architecture are not known.*

7. Conclusions

The approach outlined here may have application whenever one has several standards with which other substances can be compared for structural similarity.

This, of course, is frequently the case in medicinal and pharmaceutical chemistry, but it is evident that the technique could have broad applications in chemistry as well as in biological areas. The approach is rather economical. Even if one eventually comes across uninteresting structures, the effort involved is not necessarily wasteful, since the results can be used in detecting other undesirable structures. When the search suggests new active compounds the process can be accelerated, particularly if the candidate structure is found to exceed the initial standards in activity and thus qualifies as a standard in subsequent searches. Finally, it should be possible in an analogous manner to screen compounds for more than a single property as well as for undesirable effects (e.g., high toxicity) and combine such efforts in a single strategy.

Our approach follows from general graph theoretical foundations and strictly speaking is applicable to molecular *graphs* rather than actual molecular structures. Therefore the method uses very little chemical information (e.g., kinds of atoms, bond types, stereochemistry). The evaluated degree of similarity refers to similarity with respect to molecular *connectivity*, which is the basic concept of graph theory. This is clearly just one of the important factors necessary to consider in comparing molecules and their physical, chemical or biological properties. In the usual parlance, the assumption of similarity among structures goes beyond the simple similarity of molecular graphs. Our attempts at the present stage may be characterized as limited in scope, and aim to show how much one can deduce from such a restricted outlook on structures. Nevertheless, it is surprising how the simple concept of path codes can be utilized to provide a wealth of qualitative and quantitative information about biologically active compounds. The approach represents a new and impressively revealing application of graph theory as such, but incorporation of additional structural features can only assist and possibly further clarify our understanding of the differences which go beyond the connectivity. The inability to discriminate different kinds of atoms will restrict applications to problems where the role of functional groups appears to be critical if such groups differ in atomic composition. However, if the crucial parameters involve positional isomers and fragments belonging to the same family (e.g., alkyl) graph theoretical deductions, usually confined to combinatorial and topological traits and variations, are likely to lead to valid conclusions. In fact, some of the apparent limitations of the connectivity concept can be tackled, such as the appearance of a heteroatom or multiple bond, by extending the enumeration to such features and recording them separately [33]. If we wished we could have discriminated between the aromatic ring and the aliphatic ring of the tetralin parent structure, or we could count separately paths involving the nitrogen atom, but since these structural features appear in *all* the compounds considered we did not do so. The thrust of the present paper is not so much to discuss a particular application, but to *outline* the method, which in view of great simplicity and ease of application of the steps involved may be of interest in other similar problems. That we have arrived at the same conclusions regarding the structures to be tested that would have been obtained using the usual qualitative intuitive approach only confirms the correctness of our method.

Acknowledgment. Support of this research under Grants CHE-76-21295 and CHE-79-10263 from the National Science Foundation is gratefully acknowledged.

References

1. On leave from Ames Laboratory – DOE, Iowa State University, Ames, Iowa 50011
2. Benyon, J. H., Saunders, R. A., Williams, A. E.: The mass spectra of organic molecules, p. 121. Amsterdam: Elsevier Publishing Company 1968
3. Lang, L., Holly, S., Sohar, P.: Absorption spectra in the infrared region, pp. 102–103. London: Butterworths and Budapest: Akademia Kiado 1974
4. Prelog, V., Ružička, L.: Helv. Chim. Acta **27**, 61, 66 (1944)
5. Longuet-Higgins, H. C.: Quarterly Rev. **11**, 121 (1957)
6. Hansch, C.: Acc. Chem. Res. **2**, 232 (1969)
7. Free, S. M., Jr., Wilson, J. W.: J. Med. Chem. **7**, 395 (1964)
8. Randić, M.: J. Am. Chem. Soc. **97**, 6609 (1975)
9. Kier, L. B., Hall, L. H.: Molecular connectivity in chemistry and drug research. New York: Academic Press 1976. For a comparative analysis of molecular connectivity index, Hansch, Free–Wilson, and DARC-PELCO methods in structure-activity relationships, see: Hall, L. H., Kier, L. B.: Eur. J. Med. Chem. **13**, 89 (1978)
10. Cammarata, A., Menon, G. K.: J. Med. Chem. **19**, 739 (1976). For an outline of "pattern recognition" approaches see: Jurs, P. C., Isenhour, T. L.: Chemical applications of pattern recognition. New York: Wiley–Interscience 1975
11. Schreider, Ju. A.: Equality, resemblance, and order. Moscow: Mir Publishers 1971 (English translation, 1975)
12. Bayer, W. A., Stein, M. L., Smith, T. F., Ulam, S. M.: Math. Biosciences **19**, 9 (1974)
13. Dubois, J. E., Laurent, D., Aranda, A.: J. Chim. Phys. **70**, 1608 (1973)
14. Balaban, A. T.: Chemical applications of graph theory. New York: Academic Press 1976
15. Lieb, E. H., Mattis, D. C.: Phys. Rev. **1235**, 164 (1962); Ruch, E.: Acc. Chem. Res. **5**, 49 (1972); Gutman, I., Randić, M.: Chem. Phys. Letters **47**, 15 (1977); Randić, M.: Chem. Phys. Letters **55**, 547 (1978)
16. Muirhead, R. F.: Proc. Edinburgh Math. Soc. **21**, 144 (1903); Hardy, G. H., Littlewood, J. E., Polya, G.: Inequalities, p. 44. Cambridge Univ. Press 1934
17. Karamata, J.: Publ. Math. Univ. Belgrade **1**, 145 (1932); Beckenbach, E. F.; Bellman, R.: Inequalities. Berlin: Springer-Verlag 1961
18. Randić, M., Wilkins, C. L.: J. Chem. Info. & Computer Sci. **19**, 31 (1979)
19. For an introduction to Graph Theory see: Wilson, R. J.: Introduction to graph theory. New York: Academic Press 1972
20. Randić, M., Brissey, G. M., Spencer, R. B., Wilkins, C. L.: Computers and Chemistry **3**, 5 (1979)
21. Randić, M.: J.C.S. Faraday II, submitted
22. A *systematic* approach to molecular additivities based on graph theoretical foundations originates with Smolenskii, E. A.: Russ. J. Phys. Chem. **38**, 700 (1974). The subject has been further expanded into a model of "Graph-like State of Matter" by M. Gordon and coworkers (e.g., Gordon, M., Kennedy, J. W.: J.C.S. Faraday II. **68**, 484 (1972)). Information about more distant parts of the structure also underlies the DARC scheme: Dubois, J. E.: Entropie **21**, 5 (1968); Dubois, J. E., Laurent, D.: C.R. Acad. Sci. Paris **265C**, 780 (1967)
23. Platt, J. R.: J. Phys. Chem. **56**, 328 (1952); Altenburg, K.: Kolloid Zeit. **178**, 112 (1961)
24. Randić, M., Wilkins, C. L.: Chem. Phys. Letters **63**, 332 (1979); Randić, M., Wilkins, C. L.: J. Phys. Chem. **83**, 1525 (1979)
25. For a review on paths, their enumeration, and some properties of the derived sequences see: Randić, M.: MATCH (Inf. Commun. Math. Chem.) **7**, 5 (1979). (A plenary lecture at Bremen Conference on Mathematical Structures in Chemistry, Bremen, June 1978)
26. Interestingly, the average atom codes may reflect some salient structural properties as has been recently discerned for alkane carbon-13 chemical shift sums (Randić, M.: J. Magn. Reson., in press)

27. McDermed, J. D., McKenzie, G. M., Phillips, A. P.: J. Med. Chem. **18**, 362 (1975); Cannon, J. G., Kim, J. C., Aleem, M. A., Long, J. P.: J. Med. Chem. **15**, 348 (1972); Canon, J. G., Lee, T., Goldman, H. D.: J. Med. Chem. **20**, 1111 (1977)

28. A collection of objects defined by non-negative real numbers belongs to a *metric space* if for any pair (X, Y) a number D (called distance) can be assigned which satisfies the following:

$$D(X, Y) > 0 \quad \text{for } X \neq Y$$
$$D(X, Y) = 0 \quad \text{only if } X = Y$$
$$D(X, Y) = D(Y, X) \quad \text{for all X and Y}$$
$$D(X, Y) \leqslant D(X, Z) + D(Z, Y) \quad \text{for all X, Y, Z.}$$

For the basic definitions of similarity measures and metric spaces see: Jardine, N., Sibson, R.: Mathematical taxonomy. New York: Wiley 1971

29. The general problem of metric on sets, patterns, and coded sequences in the natural sciences is outlined in detail in: Beyer, W. A., Smith, T. F., Stein, M. L.: Metrics in biology, an introduction. Univ. of California, Los Alamos, N.M.: Los Alamos Sci. Lab. Report LA-4973, 1972

30. By "heavy" atoms, we mean all atoms except hydrogens. It is customary in applications of graph theory to molecular skeletons to suppress hydrogens (since their presence can generally be inferred)

31. Randić, M., Wilkins, C. L.: Intern. J. Quantum Chem., Quantum Biol. Symp. **6**, 55 (1979)

32. It is possible that two structures being nonisomorphic have a zero similarity index since the molecular path codes need not be unique to a structure. For example the molecular path counts for 2,3,4-trimethylhexane and 3-methyl-3-ethylhexane are identical (Ref. [24])

33. A modified program for enumeration of paths in compounds having double bonds, or bonds of different weight has been outlined: Randić, M., Brissey, G. M., Spencer, R. B., Wilkins, C. L.: Computers in Chemistry **4**, 27 (1980)

Received April 23, 1979/July 1, 1980