

characteristics make our IDB processing (classification, keywords, etc.) "human independent".

As long as technology continues to develop, the processing time at input and retrieval can be optimized, even when the actual time of retrieval is short enough: 2-4 min for a specific pattern within 100 images. Also, technology will permit one to work with a bigger storage capacity. Software development such as image compaction and image processing also will make it possible to work with the more sophisticated IDBs that are necessary in science. Further work in this direction is under way in our laboratory, where the reliability or accuracy of the IDB is being taken into account.

It is not difficult to realize that future chemical databases will allow one to work with alphanumeric, molecular, and graphic information in an easy way such as this paper has shown. Also, we can extrapolate this statement for science in general—the treatment of integrated alphanumeric and graphic information will be the normal way of managing scientific knowledge.

ACKNOWLEDGMENTS

We thank the ANDES Foundation, Grant C-10390/1988, for the Microvax II. We also thank Prof. M. L. Contreras

for discussions and FONDECYT and the University of Santiago for financial support.

REFERENCES AND NOTES

- (1) Contreras, M. L.; Deliz, M.; Galaz, A.; Rozas, R.; Sepulveda, N. A. Microcomputer-Based System for Chemical Information and Molecular Structure Search. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 105-108.
- (2) Contreras, M. L.; Deliz, M.; Rozas, R. Personal Microcomputer Based System of Chemical Information with Topological Structure Data Elaboration. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 163-167.
- (3) Rumble, J. H., Jr.; Lide, D. R., Jr. Chemical and Spectral Databases: A Look into the Future. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 231-235.
- (4) Prasad, B. E.; Gupta, A.; Toong, H.-M. D.; Madnik, S. E. A Microcomputer-Based Image Database Management System. *IEEE Trans. Ind. Electron.* **1987**, *IE-34*, 83-88.
- (5) Felician, L. Image Base Management System: a Promising Tool in the Large Office System Environment. *DATA BASE 1987/88* (Fall/Winter), 29-36.
- (6) Werman, M.; Wu, A. Y.; Melter, R. A. Recognition and Characterization of Digitized Curves. *Pattern Recognit. Lett. (Netherlands)*, **1987**, *5*, 207-213.
- (7) HP-Scanjet, Scanning Gallery Users Guide; Hewlett-Packard: Sunnyvale, CA, 1987.
- (8) Lozover, O.; Preiss, K. Automatic Construction of a Cubic B-Spline Representation for a General Curve. *Comput. Graphics* **1983**, *2*, 149-153.
- (9) Gonzalez, R. C.; Wintz, P. *Digital Image Processing*; Addison-Wesley: Reading, MA, 1977.
- (10) Kerningham, B.; Ritchie, D. *The C Programming Language*; Prentice-Hall: Englewood Cliffs, NJ, 1978.
- (11) Turbo Pascal, Reference Manual, Version 3.0; Borland International, 1985.

Vertex Indices of Molecular Graphs in Structure-Activity Relationships: A Study of the Convulsant-Anticonvulsant Activity of Barbiturates and the Carcinogenicity of Unsubstituted Polycyclic Aromatic Hydrocarbons

G. KLOPMAN* and C. RAYCHAUDHURY

Department of Chemistry, Case Western Reserve University, Cleveland, Ohio 44106

Received April 21, 1989

A new methodology is proposed whereby *local* distance based vertex indices are used in structure-activity studies. It is also shown that it is possible to reconstruct chemical graphs for those indices found to be relevant to activity. This is essential if the results of structure-activity analysis by methods utilizing graph indices are to be useful in the design of new active molecular entities. The methodology is illustrated by applications to the study of the convulsant-anticonvulsant activity of barbiturates and the carcinogenic activity of unsubstituted polycyclic aromatic hydrocarbons.

INTRODUCTION

Explaining biological activities of chemical compounds in terms of molecular topology has gained substantial attention in recent years.¹⁻¹⁸ The objective of all such studies is to explore the role of the connectedness of atoms in the expression of the biological activities of molecules. Franke et al.⁸ have discussed the necessity of considering topological aspects of the chemical structures to explain their biological functions.

The connectedness, or the topology of the molecules, is conveniently expressed in two ways. One is in terms of molecular fragments or substructures,^{1,2,5-8} and the other is in the form of molecular graphs and the indices derived therefrom.^{3,4,10-18} While substructural analyses are designed mainly to identify the potential structural components that could be responsible for some biological activity, graph-theoretical methods are mainly used to relate the structural characteristics

of chemical compounds to their biological activities. These structural characteristics include branching patterns, bonding types, cyclicity, etc.

Clearly, the substructural approaches help medicinal chemists to analyze the relationship of the molecular fragments to the biological activities of chemical compounds. However, a possible shortcoming of this approach, as well as most other approaches using discrete descriptors, is that biologically relevant substructures may be ignored if they are not present in the training data set. It appears that, topologically (in the sense of connectivity), flexible structural descriptors might play some important role in these situations. Hence, graph theory^{19,20} seems to be a prime choice to cope with such problems.

The structural formulas of chemical compounds are essentially molecular graphs whose vertices and edges represent, respectively, the atoms and their connecting chemical bonds in the molecules. This straightforward representation of chemical structures has enabled chemists for decades to take advantage of this branch of mathematics to solve some relevant

* Address correspondence to this author.

problems. From the time of Cayley,²¹ who used graph theory for the enumeration of structural isomers, until the present time when this branch of mathematics is being used to pervade the unknown world of molecular structure-biological function relationships, which has become known as structure-activity relationship (SAR) studies, the usefulness of graph theory has been quite encouraging.

In graph-theoretical structure-activity relationship (GTSAR) studies, graph invariants, in the form of real numbers known as topological indices,²²⁻²⁶ are derived from the molecular graphs. Subsequently, these indices are used as parameters in structure-activity programs.^{3,10-14,17,18} Graph-theoretical approaches have also been used in structural similarity-biological activity relationship studies.^{27,28} The underlying hypothesis in these approaches is that similar structures possess similar biological activities. However, the possibility always exists that two molecules (molecular graphs) which are structurally very close may produce considerably different biological activities as in the case of convulsant-anticonvulsant barbiturates,²⁹ which we will deal with in this paper. It thus appears that a methodology which is capable, with adequate precision, to identify even minor structural changes, such as the structural difference between two non-isomorphic graphs, is desirable for effective SAR studies.

To achieve this goal, we propose a new methodology where graph theory will be used in such a way that the difference in biological activity due to small changes in structure will be adequately taken care of. In doing this we will use *distance-based graph-theoretical indices for the vertices of the molecular graphs*. This contrasts with most other approaches where indices for the entire molecule are used. The indices will be computed only for the non-hydrogen atoms of the molecules. However, to get more discrimination of the atoms with respect to their topological environments in the respective molecules, the index values will be computed from the hydrogen-preserving molecular graphs of the compounds. The idea of using vertex indices seems quite consistent with our intention since some useful structural information of a molecule may be lost in the process of packing them into a single value. Specifically, we intend to translate the topological characteristics of a molecule with respect to each of its vertices into vertex indices, which we will refer to as local indices. We believe that a methodology utilizing a number of local indices instead of a single global index for the entire molecule would be able to extract the structural characteristics of a molecule in greater detail. Subsequently, these (local) vertex indices will be used in a newly developed methodology for the identification of topologically relevant vertices which could be related to the biological activities of the chemical compounds.

In this paper the above-mentioned SAR approach will be described and used to evaluate qualitatively the convulsant-anticonvulsant activity of oxobarbiturates and the carcinogenic activity of unsubstituted polycyclic aromatic hydrocarbons (PAHs). Two vertex indices called *vertex distance complexity* (V^d)³⁰ and *normalized vertex distance complexity* (V_n^d) will be used for our purpose. These indices, for each vertex, are computed on the basis of the topological distances of all the vertices in a molecule from that vertex by using Shannon's information formula.³¹ Using a normalized index eliminates the effect due to variations in the size of the molecules in a database since the V^d value of a vertex, in general, depends on the size of the molecule to which it belongs. A comparative study regarding the applicability of these two indices in the two databases has also been carried out.

Another problem that will be addressed in this paper concerns the construction of structures of novel analogues of some known lead compound, also referred to as lead optimization, and/or that of new lead compounds for the biological action

of interest. Structure-activity methodologies are developed to investigate the possible relationship between structural features of chemical compounds and their biological activities in terms of structural descriptors. The intent is to provide drug designers information regarding the structural requirement for an optimized or a new lead compound. So far, to our knowledge, no attempt has been made to generate structures of novel bioactive compounds on the basis of graph-theoretical information obtained from the study of a database. An attempt will be made, perhaps for the first time in the field of rational drug design, to outline an approach in this direction using graph reconstruction techniques. The role of distance-based vertex indices for such purpose will also be discussed.

COMPUTATION OF VERTEX INDICES

Unless specified, a graph G will mean an undirected, unweighted graph whose pairs of vertices are connected by single edges, and there is no self-loop on any vertex of the graph. If u and v are two vertices of G , then a path $p(uv)$ between u and v in G is defined as an alternate sequence of vertices and edges of the form $ue_1u_1e_2u_2...u_{l-1}e_lv$, where the vertices $u, u_1, u_2, ..., u_{l-1}, v$ and the edges $e_1, e_2, ..., e_l$ are distinct. The path $p(uv)$ is said to be of length l since it contains l edges. The distance $d(u, v)$ between u and v is the length of the shortest path between u and v .

Let G have n vertices. If there are n_k vertices at a distance d_k , where $k = 1, 2, ..., m$, from any vertex v in G , then let

$$d(v) = \sum_{k=1}^m n_k d_k \quad (1)$$

where d_m is the distance of the most distant vertex (vertices) from v in the graph G . Let us now consider that each d_k is the component of a partition of $d(v)$. On the basis of this partition a measure of information content, called "vertex distance complexity (V^d)" of v can be obtained³⁰ by using Shannon's formula.³¹ The V^d value of v , $V^d(v)$, is given by

$$\begin{aligned} V^d(v) &= - \sum_{k=1}^m n_k (d_k/d(v)) \text{lb} (d_k/d(v)) \\ &= \sum_{k=1}^m n_k (d_k/d(v)) \text{lb} (d(v)/d_k) \end{aligned} \quad (2)$$

where lb stands for \log_2 and the index V^d is expressed in bits.

The normalization of the index V^d can be done by dividing the V^d value of a vertex by $\text{lb } d$, the maximum V^d value that can be obtained from a partition of $d(v)$. Thus, normalized vertex distance complexity, V_n^d , may be computed by using

$$V_n^d(v) = V^d(v) / \text{lb } d(v) \quad (3)$$

It may be noted that the d_k values for a vertex may be obtained from the entries of the corresponding row/column of the distance matrix $D(G)$ of G . This makes the computation of the distance-based vertex indices on a computer very convenient. The distance matrix $D(G)$ of a graph G that has n vertices is defined as

$$D_{ii} = 0 \quad D_{ij} = d(v_i, v_j) \quad (4)$$

where D_{ij} is the (i, j) th entry in $D(G)$ and gives the distance between two vertices v_i and v_j in G , $i, j = 1, 2, ..., n$. The computation of V^d from a distance matrix is illustrated below for the hydrogen-preserving molecular graph of benzene.

The molecular graph G' and the corresponding distance matrix $D(G')$ of benzene are given in Figure 1. We will now illustrate the computation of the V^d and V_n^d values of the vertices labeled 1-6 in G' since these vertices represent the non-hydrogen atoms in that molecule. Considering the vertex 1 (v_1) of G' , we see from the distance matrix $D(G')$ that there are three vertices at a distance 1, four vertices at a distance

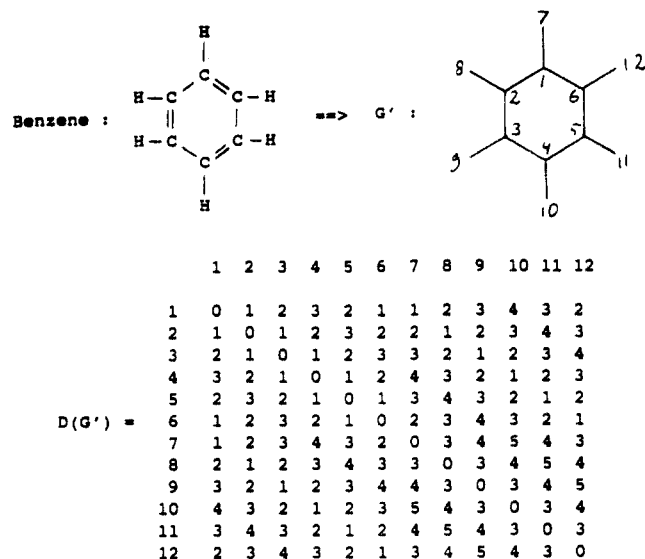


Figure 1. Hydrogen-preserving molecular graph G' and the corresponding distance matrix $D(G')$ of benzene.

2, three vertices at a distance 3, and one vertex at a distance 4 from v_1 in G' . From these distance values we find, using (1), that

$$d(v_1) = (3 \times 1) + (4 \times 2) + (3 \times 3) + (1 \times 4) = 24.$$

Now, considering 24(1,1,1,2,2,2,3,3,4) as a partition of 24, the V^d and the V_n^d values of v_1 can be calculated by using (2) and (3). Thus

$$V^d(v_1) = 3(1/24) \lg(24/1) + 4(2/24) \lg(24/2) + 3(3/24) \lg(24/3) + 1(4/24) \lg(24/4) = 3.3239 \text{ bits}$$

$$V_n^d(v_1) = 3.3239 / \lg 24 = 0.7250$$

Evidently, V^d and V_n^d values of vertices 1–6 are the same.

METHODOLOGY

The basic premise of the present approach is to characterize each atom of the compounds of a database with respect to its topological position in the respective molecular environments such that this structural information may be used to segregate active from inactive compounds. The structural information will be depicted in the form of real numbered vertex indices which would be expected to take care of the topological positions of the atoms of a molecule relative to other atoms in that molecule in a fairly discriminating manner. In this analysis, no attempt has been made to account for different atom or bond types, although, in principle, this can be done by using other graph indices. Our purpose here is to establish the principles of the new methodology which, if warranted, can be refined at a later stage.

Once a training set of active and inactive compounds is assembled, the molecular structures of the compounds are entered in a computer program via the KLN code³² or a graphical input. The program computes the index values of the non-hydrogen atoms of all the compounds in the database. The computed values are then arranged in a sequence of nonincreasing (or nondecreasing) order, and ranges of values contributed by active and inactive molecules are identified in the ordering. A range composed of values coming from active molecules will be called an *active range*. Similarly, an *inactive range* will be composed of values coming from inactive molecules. It is believed that these ranges contain those vertices whose topological positions distinguish active compounds from the inactive ones. Hence, these ranges may be used to identify active and inactive compounds. For example, a compound may

be predicted active if the index values of all or some of its vertices fall in active range(s). Presumably, the consideration of vertex indices, i.e., a number of local indices for each molecule, makes it possible to obtain various active and inactive ranges. These ranges, in turn, are expected to help judge and identify active molecules with high accuracy on the basis of the occurrence of the index values of their atoms in these ranges. Certain criteria have been adopted for the selection of ranges in an ordering and for the prediction of activity. We describe below the way our computer program performs it.

Once the vertex index values for all the compounds in the database are arranged in the order mentioned above, the computer program starts searching active and inactive ranges from the lowest value in the ordering. If the lowest value comes from, say, an active molecule, then the program will verify whether the next higher value is also from an active molecule. The program continues this procedure until it finds a value coming from an inactive molecule. The program will qualify the values coming from the active molecules to form an active range if certain criteria are satisfied. The same procedure is carried out over the entire ordering to identify different active and inactive ranges.

The criteria for the selection of ranges are as follows: (a) A value coming more than once from the same molecule is only considered once. This is so because topologically equivalent vertices in a graph will have the same index value. (b) Three or more consecutive values coming exclusively from active molecules or exclusively from inactive molecules are considered to form an active range or an inactive range, respectively, if at least three of them are distinct when coming from the same molecule or at least two of them are distinct when coming from more than one molecule. (c) Some single value coming from both active and inactive molecules is not considered to form, say, an active range by itself or together with other value(s) unless more than two-thirds of the molecules contributing this value are active. The same rule is applicable regarding an inactive range as well.

The ordering may also contain several regions where active or inactive ranges do not occur. The formation of ranges in an ordering is illustrated in Figure 2 considering the V_n^d values obtained for the training set of the PAH database shown in Table II.

Figure 2 shows that the values corresponding to serial numbers 24–27 do not belong to a range since the single value 0.6434 has been contributed by two active and one inactive compound, although three of those four values are coming from active molecules. The inactive range (34–38) of five values has been formed by three distinct values coming from five different inactive compounds, and the active range (48–54) of seven values contains six distinct values originating in six compounds.

The activity of a compound is predicted on the basis of the occurrence of the vertex index values of a molecule in active and inactive ranges. A compound is predicted to be active if the index values of all or most of its atoms fall (a) only in the active range(s) or (b) in both active and inactive ranges and the number of index values falling within the active range(s) is larger than that falling in the inactive range(s). Otherwise, the compound is predicted to be inactive. The activity prediction of the second compound in the barbiturate test set (Table I) on the basis of V_n^d values of its atoms is illustrated in Figure 3.

SAR RESULTS

In the present study two information-theoretical vertex indices, viz., vertex distance complexity (V^d)³⁰ and normalized vertex distance complexity (V_n^d), have been used to evaluate qualitatively the convulsant–anticonvulsant activity of oxo-

Serial No.*	V_n^d value	Compound No.	Activity of the compound	Type of range
24	0.6434	4	-	not a range
25	0.6434	6	+	
26	0.6434	7	+	
27	0.6435	23	+	
34	0.6484	10	-	Inactive
35	0.6484	11	-	
36	0.6484	14	-	
37	0.6486	24	-	
38	0.6491	28	-	
48	0.6519	7	+	Active
49	0.6527	23	+	
50	0.6531	12	+	
51	0.6531	16	+	
52	0.6535	23	+	
53	0.6545	20	+	
54	0.6547	21	+	

* The serial no. corresponds to the ordering of the V_n^d values.

Figure 2. Sample of one active and one inactive range along with a situation where a range has not occurred. Column 4 gives activity of the compounds in that row (+ = active; - = inactive).

Serial No.	V_n^d value	Falling in the range	Number of values in the range from : actives	inactives
1.	0.6829	Not falling in any range		
2.	0.7114	"		
3.	0.7443	"		
4.	0.6850	"		
5.	0.7125	"		
6.	0.6860	"		
7.	0.6477	"		
8.	0.6639	0.6635 - 0.6646	3	0
9.	0.6860	Not falling in any range		
10.	0.6850	"		
11.	0.7125	"		
12.	0.7485	0.7481 - 0.7495	0	3
13.	0.7097	Not falling in any range		
14.	0.7298	0.7298 - 0.7326	3	0
15.	0.7015	Not falling in any range		
16.	0.6756	0.6753 - 0.6787	4	0
17.	0.6756	0.6753 - 0.6787	4	0

PREDICTION : The molecule is ACTIVE (Convulsant).

Figure 3. Activity prediction of the second compound in the barbiturate test set (Table I) using the normalized index V_n^d .

barbiturates and the carcinogenic activity of unsubstituted polycyclic aromatic hydrocarbons (PAHs). A comparative study has also been carried out by using these two indices to investigate the predictive capacity of these indices for each database. The data for the barbiturates²⁹ and the PAHs³³ have been collected from the literature. Tables I and II, respectively, show the qualitative evaluation of the convulsant-anticonvulsant activity of barbiturates and the carcinogenic activity of the PAHs using the normalized index V_n^d .

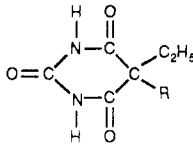
For the barbiturates a training set of 15 compounds and a test set of 4 compounds were created by a random choice of compounds from a database of 19 oxobarbiturates²⁹ considering convulsant barbiturates to be active and the anticonvulsants to be inactive. The criterion for anticonvulsant action of the barbiturates is prevention of the tonic extensor seizure produced by pentylenetetrazole. It is clear from the results furnished in Table I that all the compounds of both the training and the test sets have been classified correctly on the basis of the normalized index V_n^d . When these compounds were evaluated by using the index V^d , all the compounds of the training set were again classified correctly. However, in this case, the classification of the test-set compounds was extremely

poor since three of the four compounds were mispredicted (Table I).

The PAH database consisted of the 41 unsubstituted polycyclic aromatic hydrocarbons listed by Dipple.³³ The results of the evaluation of their carcinogenic activity are given in Table II. The 41-compound data set was divided in two groups consisting of a training set of 28 compounds and a test set of 13 compounds. The compounds were listed as inactive if they are so labeled in the reference³³ and as active if they are labeled moderately or very active. For this database, V_n^d produced 2 misclassifications of the 28 compounds in the training set and 2 misclassifications of the 13 compounds in the test set. In this database the index V^d produced only slightly inferior results for the test set where only 3 of the 13 compounds were misclassified. In the training set there was only one incorrect prediction.

It may be recalled that the normalized index was introduced in this study with a view to eliminate the effect due to large variations in the size of the molecules. Clearly, in both databases, the barbiturates and the PAHs, there are wide variations in the size of the molecules. It is therefore interesting to note that the normalized index has indeed produced better

Table I. Qualitative Evaluation of the Convulsant–Anticonvulsant Activity of Barbiturates

serial no.		activity		
		assigned ^a	V_n^d	V^d
Training Set				
1	R = $-(CH_2)_3CH_3$	+	+	+
2	$-CH(CH_3)(CH_2)_2CH_3$	-	-	-
3	$-(CH_2)_2CH(CH_3)_2$	-	-	-
4	$-CH(CH_3)CH_2CH(CH_3)_2$	-	-	-
5	$-CH=CHCH_2CH_3$	-	-	-
6	$-C(CH_3)=CHCH_2CH_3$	-	-	-
7	$-CH_2CH=CHCH_3$	-	-	-
8	$-CH(CH_3)CH=CHCH_3$	-	-	-
9	$-CH_2CH=C(CH_3)_2$	+	+	+
10	$-CH(CH_3)CH=C(CH_3)_2$	+	+	+
11	$-(CH_2)_3C_6H_{11}$	-	-	-
12	$-(CH_2)_2CH=C_6H_{10}$	+	+	+
13	$-(CH_2)_2CH=C_5H_8$	+	+	+
14	$-CH_2C_6H_5$	-	-	-
15	$-CH_2CH(CH_3)C_6H_5$	+	+	+
Test Set				
1	R = $-CH=(CH)_2(CH_3)_2$	-	-	-
2	$-C(CH_3)=(CH)_2(CH_3)_2$	+	+	-
3	$-(CH_2)_3C_6H_5$	-	-	+
4	$-(CH_2)_2C_6H_5$	+	+	-

^a The data have been taken from Andrews et al.²⁹ The anticonvulsant activity of the barbiturates was tested against pentylenetetrazole.²⁹ The convulsants represent actives (+), and the anticonvulsants represent inactives (-).

result for both databases. In our previous work with the mutagenic activity of substituted nonfused ring aromatic compounds¹⁷ and the hallucinogenic activity of substituted phenylalkylamines,¹⁸ the V^d index produced better results (though no result of comparative study was reported in those papers). In both of those databases the compounds were congeneric. Thus, it appears that the normalized index yields better results in molecules of widely varied structural size, while V^d is possibly more effective for congeneric databases. However, studies with more databases would, perhaps, reveal the truer picture regarding the usefulness of a particular index in a database of interest.

To investigate whether the ranges obtained in the orderings using the normalized index V_n^d are related to the activity, we performed a number of analyses on scrambled sets. To do this, we randomly reallocated the observed activities to the molecules of the database and analyzed the resultant scrambled sets as if they were experimentally observed. We performed this operation three times each for the barbiturates and for the PAHs and evaluated the data for both types of compounds on the basis of the ranges obtained each time. In every one of these cases, the observed ranges provided adequate reclassification of the training set. This indicates that the attainment of a good reclassification of the training set is not a good indicator of meaningful structure–activity relationships. On the other hand, the ability to make predictions on the basis of these ranges deteriorated significantly. Indeed, it was found that for the barbiturates there were two misclassifications of the four compounds of the test set in all of the three experiments with the scrambled data. For the PAHs there were five misclassifications of the test-set compounds in two occasions and six misclassifications in one occasion. It is apparent from these studies that 2 mispredictions of 4 compounds for the barbiturates and 5 or 6 mispredictions of 13 compounds for the PAHs may be regarded as outcomes that could happen by chance. Hence, it appears that the evaluations based on the real data are far from chance predictions, and therefore the observed ranges may be regarded as relevant to activity.

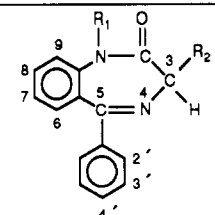
Table II. Qualitative Evaluation of Carcinogenicity of Unsubstituted Polycyclic Aromatic Hydrocarbons

serial no.	compound name	activity		
		assigned*	V_n^d	V^d
Training Set				
1	benzene	-	-	-
2	naphthalene	-	-	-
3	phenanthrene	-	-	-
4	naphthacene	-	-	-
5	triphenylene	-	-	-
6	chrysene	+	+	+
7	benz[a]anthracene	+	+	+
8	benz[e]pyrene	-	-	-
9	perylene	-	-	-
10	benz[a]naphthacene	-	-	-
11	benz[b]chrysene	-	-	-
12	benz[c]chrysene	+	+	+
13	benz[g]chrysene	+	+	+
14	picene	-	-	-
15	dibenz[ac]anthracene	+	+	+
16	dibenz[aj]anthracene	+	+	+
17	dibenz[cg]phenanthrene	-	+	+
18	anthanthrene	-	-	-
19	benz[ghi]perylene	+	+	+
20	dibenz[al]pyrene	+	+	+
21	dibenz[ai]pyrene	+	+	+
22	dibenz[el]pyrene	-	-	-
23	naphtho[2,3-a]pyrene	+	+	+
24	naphtho[2,3-e]pyrene	-	-	-
25	dibenz[aj]naphthacene	-	-	-
26	dibenz[ac]naphthacene	+	-	-
27	anth[1,2-a]anthracene	-	-	-
28	benz[b]pentaphene	-	-	-
Test Set				
1	anthracene	-	-	-
2	pyrene	-	-	-
3	benz[c]phenanthrene	+	+	+
4	benz[a]pyrene	+	+	-
5	pentacene	-	-	-
6	dibenz[ah]anthracene	+	-	-
7	dibenz[bq]phenanthrene	-	+	+
8	pentaphene	-	-	-
9	dibenz[ae]pyrene	+	+	+
10	dibenz[ah]pyrene	+	+	+
11	dibenz[bk]chrysene	-	-	-
12	benz[c]pentaphene	-	-	-
13	naphtho[1,2-a]triphenylene	-	-	-

* The data were collected from Dipple.³³ The carcinogens are actives (+), and the noncarcinogens are inactives (-).

We were also interested in investigating whether the ranges obtained from the classification of convulsant–anticonvulsant barbiturates would be useful to evaluate other types of anti-convulsant agents. Since very minor structural changes turn a convulsant barbiturate into an anticonvulsant one, we suspect that the topological requirements for such activities, convulsant activity in particular, may be very tight. We found a series of 25 benzodiazepines³⁴ (Table III) whose anticonvulsant activity had also been tested against pentylenetetrazole, as were the barbiturates, and tested them on the basis of the ranges obtained from the barbiturate database. What emerged from this experiment was that 23 of the 25 compounds were correctly classified as inactives, i.e., anticonvulsants. It is noted that none of the inactives (anticonvulsants) were predicted by default. Since both the convulsant and anticonvulsant properties of the molecules were predicted on the basis of the occurrence of their vertices in ranges, this finding seems to support the validity of the topological ranges obtained from the classification of the convulsant and anticonvulsant barbiturates. Furthermore, our results indicate the possibility that some underlying topological basis may exist for the convulsant–anticonvulsant activity of chemical compounds in general. Further studies with a wider variety of convulsant and anti-

Table III. Qualitative Activity Prediction of a Series of 1,4-Benzodiazepines on the Barbiturate Database

serial no.		activity			
		log (1/c) ^a	assigned ^a	pre- dicted V_n^d	V^d
1	1,4-benzodiazepine-2-ones	-0.53	-	-	-
2	7-F	-0.50	-	-	-
3	7-Cl	1.65	-	-	-
4	7-CN	2.30	-	-	+
5	7-NO ₂	2.60	-	-	-
6	7-CF ₃	2.48	-	-	-
7	7-CH ₃	0.16	-	-	-
8	7-SCH ₃	1.15	-	-	-
9	7-SO ₂ CH ₃	-0.28	-	-	+
10	7-phenyl	-0.41	-	-	+
11	8-Cl	-0.09	-	-	-
12	7-NO ₂ ; 9-CH ₃	1.56	-	-	-
13	7,8-CH ₃	0.30	-	-	-
14	7-Cl; 2'-Br	2.76	-	-	-
15	7-Cl; 2'-OCH ₃	1.60	-	-	-
16	7-CN; 2'-F	2.63	-	-	-
17	7-NO ₂ ; 2'-Cl	3.30	-	-	+
18	7,2'-NO ₂	2.97	-	-	-
19	1-CH ₃ ; 7-NO ₂	2.63	-	-	-
20	1-CH ₃ ; 7-N(CH ₃) ₂	1.69	-	+	-
21	1-CH ₃ ; 7-NO ₂ ; 2'-Cl	3.92	-	+	+
22	1-CH ₃ ; 7-N(CH ₃) ₂ ; 2'-Cl	2.82	-	+	-
23	1-CH ₃ ; 3-OH; 7-Cl	2.63	-	-	+
24	5-pyridyl; 7-Br	2.66	-	-	-
25	1-CH ₃ ; 7,2'-Cl	2.68	-	-	-

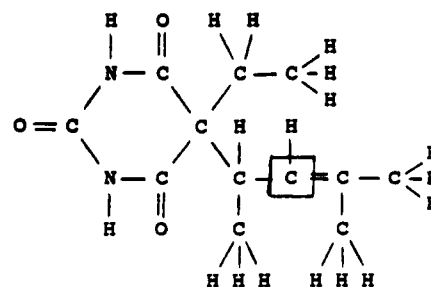
^a The data have been taken from Camerman et al.³⁴ The anticonvulsion activity is expressed by log (1/C), where C is the dose in millimoles per kilogram required to suppress completely the seizures induced by the administration of 125 mg/kg metrazole in 50% of the treated animals. The convulsants are active (+), and the anticonvulsants are inactive (-).

convulsant compounds seem worth exploring to extend the validity of this hypothesis.

We also evaluated the benzodiazepine compounds with the barbiturate database using the V_n^d index and found six mispredictions. This further documents the need to use a normalized index for diverse databases.

DESIGN OF NOVEL STRUCTURES

In this section we propose a possible algorithm for generating novel structures from the distance distribution associated with a vertex in a molecular graph. Our intent is to use graph reconstruction techniques for generating structures, given the number of vertices (atoms) to lie at different distances from a root vertex. To do this we must find the number of atoms located at different distances from each atom in a molecule. This can be done by recalling the entries of the row corresponding to the atom in the distance matrix of the molecule (molecular graph) to which the atom belongs. One can now construct different graphs where the number of vertices at different distances from a root vertex will be the same as the distance distribution associated with the vertex. This will generate a number of rooted trees. Now, if one joins the vertices situated at the same distance or at consecutive distances, then one obtains more graphs containing cycle(s) in which the number of vertices at different distances from the root remains the same. It is interesting that this approach can be used to generate new lead compounds or novel analogues of a series. We show below how such an approach may be used to generate the structure of a compound from the distance distribution associated with a vertex (atom) of a different compound. The molecule shown in Figure 4 is compound 10


Figure 4. Barbiturate: compound 10 in Table I.

of the barbiturate training set of Table I. The atom shown in the box has its V_n^d value of 0.7298 obtained from the distance distribution

1,1,1,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,5,5,5,5,5,6

associated with the atom.

Now, considering the above distance distribution, one can construct a number of rooted trees having 34 vertices lying at the given distances from a root vertex. The construction of one of the rooted trees is illustrated in Figure 5.

The molecule constructed in Figure 6 happens to be compound 2 of the barbiturate test set given in Table I. It is noted that vertex v , picked up for the construction, was found in an active range (0.7298–0.7326) of three values. Furthermore, both compound 10 of the training set and compound 2 of the test set are active, i.e., convulsant barbiturates.

Hence, the present methodology enabled us to identify useful vertices from active ranges for the construction of a new structure. Using the distance distributions associated with the atoms of known molecules, one may thus be able to predict structures of new compounds that may have the desired biological activity. As mentioned earlier, one can always construct a number of tree and cyclic graphs from a given distance distribution. Hence, one may design as many compounds as needed using one's own experience regarding the appropriate use of chemical elements in the place of the vertices in the reconstructed graphs. Now these newly designed compounds may be evaluated as a test set, and some of them may surface as useful as in our example, i.e., the reconstructed barbiturate would have been rightly predicted as active.

Again, besides considering the distance distribution associated with a vertex in an active range for the reconstruction, it may also be possible to find the distance distributions and the corresponding vertex index values that might fall in any of the active ranges in an ordering even if that value does not match any of the values in that range. Subsequently, these distance distributions may be used for graph reconstruction. In fact, such a situation has occurred in our experiment. For our earlier experiment we considered convulsant barbiturates to be active and anticonvulsants to be inactive. However, since both convulsants and anticonvulsants have been predicted on the basis of the occurrence of their vertex index values in active and inactive ranges, following the criteria for the prediction of activity, one would obtain the same correct predictions if anticonvulsants were taken as active and convulsants as inactive. Under this experimental makeup 23 of the 25 benzodiazepines would be predicted to be active, i.e., anticonvulsant. Considering anticonvulsants to be active, it has been found that atom 12 of compound 5 in Table III belongs to an active range of six values, although the index value of that vertex did not match any value in that range. Clearly, one can reconstruct the benzodiazepine compound from the distance distribution associated with that vertex. To be more precise, compound 5 in Table III would be one of the compounds which could be generated from that distance distribution and that compound would rightly be predicted as active

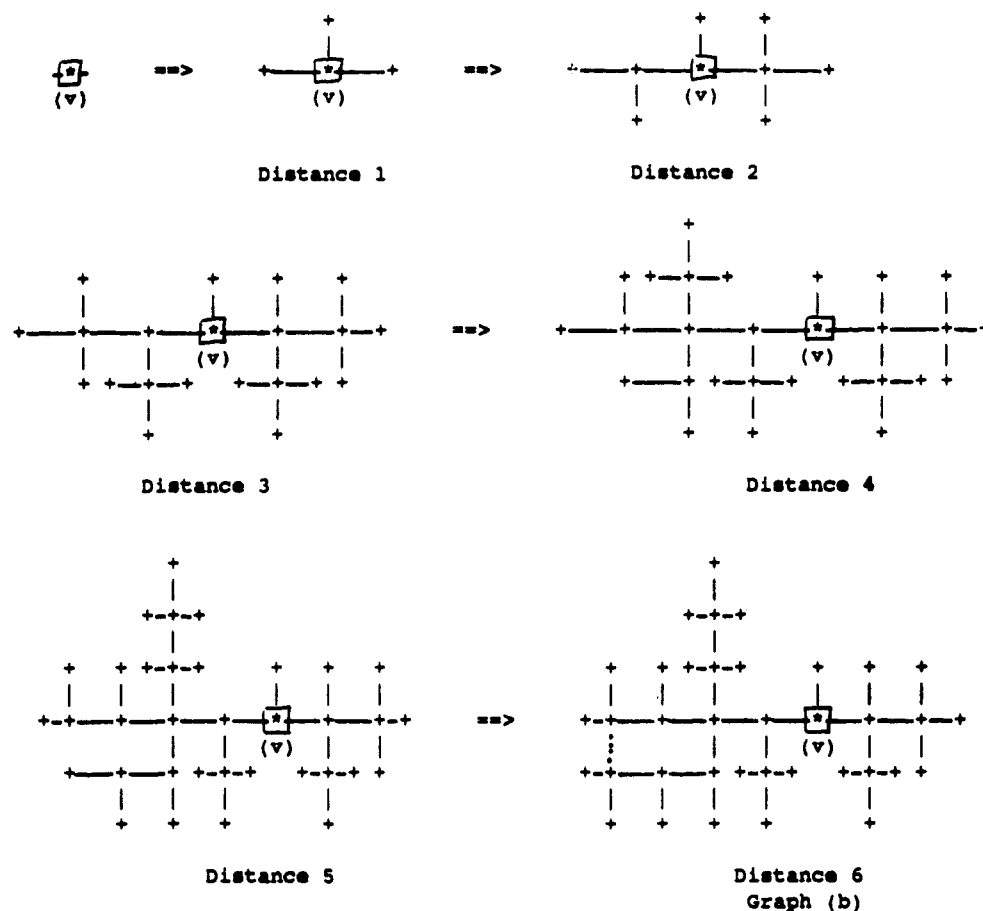


Figure 5. Rooted tree of atom $*(v)$ of compound 10 of Table I. The dots in distance graph 6 show cyclization.

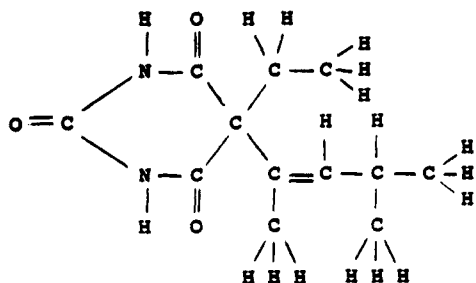


Figure 6. Reconstructed barbiturate structure (from graph b of Figure 5).

(anticonvulsant), as it has been in our experiment. This emphasizes the importance of finding possible distance distributions that may produce vertex index values to lie in an active range. This will open the possibility of generating more compounds for some biological activity of interest.

It is apparent that the number of graphs that can be constructed from a given distance distribution will increase rapidly with the increase in the number of vertices to be situated at different distances from the root vertex. However, different restrictions may be imposed on the construction of graphs so that only chemically meaningful structures are generated. For example, one of the restrictions might be that not more than three vertices should be attached to any vertex during the reconstruction. This seems to be a meaningful restriction from a chemical point of view since in most structures there are at most four first-neighbor atoms. Such restrictions are expected to reduce the number of possible graphs to a large extent. Moreover, as better knowledge is gathered regarding the applicability of this method, it may be possible to identify an appropriate vertex from a suitable active range for the purpose of generating novel structures. This will further reduce the number of graphs to be reconstructed.

ACKNOWLEDGMENT

Support by the office of Naval Research through its Selected Research Opportunities Program (N00014-84-K-0090) is highly appreciated.

Registry No. 1 (Barbiturate training set), 77-28-1; 1 (barbiturate test set), 66968-52-3; 1 (polycyclic training set), 71-43-2; 1 (polycyclic test set), 120-12-7; 1 (benzodiazepine), 2898-08-0; 2 (barbiturate training set), 76-74-4; 2 (barbiturate test set), 72961-79-6; 2 (polycyclic training set), 91-20-3; 2 (polycyclic test set), 129-00-0; 2 (benzodiazepine), 2648-00-2; 3 (barbiturate training set), 57-43-2; 3 (barbiturate test set), 124381-05-1; 3 (polycyclic training set), 85-01-8; 3 (polycyclic test set), 195-19-7; 3 (benzodiazepine), 1088-11-5; 4 (barbiturate training set), 2964-06-9; 4 (barbiturate test set), 17013-38-6; 4 (polycyclic training set), 92-24-0; 4 (polycyclic test set), 50-32-8; 4 (benzodiazepine), 17562-53-7; 5 (barbiturate training set), 2237-92-5; 5 (polycyclic training set), 217-59-4; 5 (polycyclic test set), 135-48-8; 5 (benzodiazepine), 146-22-5; 6 (barbiturate training set), 125-42-8; 6 (polycyclic training set), 218-01-9; 6 (polycyclic test set), 53-70-3; 6 (benzodiazepine), 2285-16-7; 7 (barbiturate training set), 1952-67-6; 7 (polycyclic training set), 56-55-3; 7 (polycyclic test set), 195-06-2; 7 (benzodiazepine), 5571-63-1; 8 (barbiturate training set), 17013-35-3; 8 (polycyclic training set), 192-97-2; 8 (polycyclic test set), 222-93-5; 8 (benzodiazepine), 2891-12-5; 9 (barbiturate training set), 21149-88-2; 9 (polycyclic training set), 198-55-0; 9 (polycyclic test set), 192-65-4; 9 (benzodiazepine), 6404-87-1; 10 (barbiturate training set), 3625-18-1; 10 (polycyclic training set), 226-88-0; 10 (polycyclic test set), 189-64-0; 10 (benzodiazepine), 70740-89-5; 11 (barbiturate training set), 124381-01-7; 11 (polycyclic training set), 214-17-5; 11 (polycyclic test set), 217-54-9; 11 (benzodiazepine), 5571-50-6; 12 (barbiturate training set), 124381-02-8; 12 (polycyclic training set), 194-69-4; 12 (polycyclic test set), 222-54-8; 12 (benzodiazepine), 4941-45-1; 13 (barbiturate training set), 124381-03-9; 13 (polycyclic training set), 196-78-1; 13 (polycyclic test set), 53156-66-4; 13 (benzodiazepine), 5571-57-3; 14 (barbiturate training set), 36226-64-9; 14 (polycyclic training set), 213-46-7; 14 (benzodiazepine), 63574-83-4; 15 (barbiturate training set), 124381-04-0; 15 (polycyclic training set), 215-58-7; 15 (benzodiazepine), 3023-44-7;

16 (polycyclic training set), 224-41-9; 16 (benzodiazepine), 846-58-2; 17 (polycyclic training set), 188-52-3; 17 (benzodiazepine), 1622-61-3; 18 (polycyclic training set), 191-26-4; 18 (benzodiazepine), 4980-73-8; 19 (polycyclic training set), 191-24-2; 19 (benzodiazepine), 2011-67-8; 20 (polycyclic training set), 191-30-0; 20 (benzodiazepine), 2891-09-0; 21 (polycyclic training set), 189-55-9; 21 (benzodiazepine), 5527-71-9; 22 (polycyclic training set), 192-51-8; 22 (benzodiazepine), 30144-75-3; 23 (polycyclic training set), 196-42-9; 23 (benzodiazepine), 846-50-4; 24 (polycyclic training set), 193-09-9; 24 (benzodiazepine), 1812-30-2; 25 (polycyclic training set), 227-04-3; 25 (benzodiazepine), 2894-68-0; 26 (polycyclic training set), 216-00-2; 27 (polycyclic training set), 195-00-6; 28 (polycyclic training set), 222-78-6.

REFERENCES AND NOTES

- (1) Klopman, G. *J. Am. Chem. Soc.* **1984**, *106*, 7315.
- (2) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
- (3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley-Research Studies Press, Letchworth, U.K., 1986.
- (4) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Wiley-Research Studies Press, Chichester, U.K., 1983.
- (5) Chu, K. C.; Feldman, R. J.; Shapiro, M. B.; Hazard, G. F., Jr.; Geran, R. I. *J. Med. Chem.* **1975**, *18*, 539.
- (6) Free, S. M., Jr.; Wilson, J. M. *J. Med. Chem.* **1964**, *7*, 395.
- (7) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. J. *J. Med. Chem.* **1977**, *20*, 496.
- (8) Franke, R.; Huebel, S.; Streich, W. J. *EHP, Environ. Health Perspect.* **1985**, *61*, 239.
- (9) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
- (10) Grossman, S. C.; Dzonova, B. J.-B.; Randic, M. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1986**, *12*, 123.
- (11) Raychaudhury, C.; Basak, S. C.; Roy, A. B.; Ghosh, J. J. *Indian Drugs* **1980**, *18*, 97.
- (12) Ray, S. K.; Basak, S. C.; Raychaudhury, C.; Roy, A. B.; Ghosh, J. J. *Arzneim.-Forsch./Drug Res.* **1983**, *33*(1), 501.
- (13) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *J. Pharm. Sci.* **1984**, *73*, 429.
- (14) Roy, A. B.; Raychaudhury, C.; Ghosh, J. J.; Ray, S. K.; Basak, S. C. In *Quantitative Approaches to Drug Design*; Dearden, J. C., Ed.; Elsevier: Amsterdam, 1983; p 75.
- (15) Rouvray, D. H. *Acta Pharm. Jugosl.* **1986**, *36*, 239.
- (16) Basak, S. C. *Med. Sci. Res.* (in press).
- (17) Klopman, G.; Raychaudhury, C. *J. Comput. Chem.* **1988**, *9*, 232.
- (18) Klopman, G.; Raychaudhury, C.; Henderson, R. V. *Math. Comput. Modell.* **1988**, *11*, 635.
- (19) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1972.
- (20) Deo, N. *Graph Theory with Applications to Engineering and Computer Science*; Prentice-Hall: Englewood Cliffs, NJ, 1974.
- (21) Cayley, A. *Philos. Mag.* **1874**, *47*, 444.
- (22) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. 2, Chapter 4.
- (23) Sarkar, R.; Roy, A. B.; Sarkar, P. K. *Math. Biosci.* **1978**, *39*, 299.
- (24) Basak, S. C.; Roy, A. B.; Ghosh, J. J. *Proceedings of the Second International Conference on Mathematical Modelling*; University of Missouri: Rolla, 1979; Vol. 2, p 851.
- (25) Raychaudhury, C.; Ghosh, J. J. *Proceedings of the Third Annual Conference of the Indian Society for Theory of Probability and its Applications*; Aug. 22-24, 1981; Wiley Eastern Limited: New Delhi, 1984.
- (26) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581.
- (27) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Applied Mathematics* (in press).
- (28) Johnson, M. In *Graph Theory with Applications to Algorithms and Computer Science*; Alavi, Y., Chartrand, G., Lesniak, L., Lick, D. R., Wall, C. E., Eds.; Wiley-Interscience: New York, 1985; p 457.
- (29) Andrews, P. R.; Mark, L. C.; Winkler, D. A.; Jones, G. P. *J. Med. Chem.* **1983**, *26*, 1223.
- (30) Raychaudhury, C. Ph.D. Thesis, Jadavpur University, Calcutta, India, 1983.
- (31) Shannon, C.; Weaver, W. *Mathematical Theory of Communication*; University of Illinois Press: Urbana, 1949.
- (32) Klopman, G.; McGonigal, M. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 48.
- (33) Dipple, A. *ACS Monogr.* **1984**, *182*, 41.
- (34) Camerman, A.; Camerman, N. *Acta Crystallogr.* **1981**, *B37*, 1677.

Use of Vector Processing To Search the Cambridge Structural Database

A. H. M. THIERS and J. H. NOORDIK*

CAOS/CAMM Center, Faculty of Science, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands

J. BOERHOUT

Convex Computer, Europalaan 514, 3526 KS Utrecht, The Netherlands

Received June 7, 1989

The Cambridge Structural Database (CSD) is a vast numerical resource of crystallographic data. The January 1989 release contains over 70 000 entries, and the data acquisition rate currently increases about 15% per annum. To be able to provide adequate response times for interactive data retrieval, using the new (1988) CSD file format, a vectorized search procedure has been developed as a modification of the CSD program QUEST. This procedure employs the pipelined vector facilities of the CONVEX C120 system to perform bitscreen logic, resulting in response times for arbitrary queries in the order of seconds, almost independent of the size of the database.

INTRODUCTION

The Cambridge Structural Database¹ (CSD) represents a rapidly increasing reservoir of coordinate-based information on molecular structures. This reservoir is accessible via search queries, composed primarily in chemical terms, and in different places¹⁻³ innovative use of modern computer systems and network facilities helps to deliver this information to the desk of the research chemist. Until 1988 the release files of CSD consisted of three separate files: a BIBliographic file, a CONNectivity file, and a DATA file. Only the former two were searchable with Cambridge Crystallographic Datacentre software, which originated in the 1960s. The computational slowness of this system, combined with more sophisticated user needs, necessitated a major software development effort which

resulted in the release, by the Cambridge Crystallographic Datacentre, of the CSD version 3 system in 1988 and the CSD version 4 system in 1989. The main differences with the previous system were a new unified release file and the use of bitscreens to rapidly select candidate structures for exact matches. In the bitscreens a large variety of structural, experimental, and reliability flags are set, and screen out percentages of 95% and higher are easily reached on the more common queries. This screening procedure, combined with direct access file reading, resulted in a typical CPU time requirement of about 200 s on a VAX 11/785 system for a common connectivity query against the 1989 file. On heavily loaded computer systems, this CPU requirement can make on-line searching rather impractical, and this situation will only