Rotadex—A New Index for Generic Searching of Chemical Compounds*

By IRVING H. SHER, JOHN O'CONNOR, and EUGENE GARFIELD
Institute for Scientific Information, Philadelphia, Pennsylvania

Received July 29, 1963

Many different and interesting systems have been devised for the coding of chemical compounds.[1-23] Chemical codes which do not attempt to describe compounds completely and uniquely may be termed generic codes. These codes deliberately sacrifice specificity for convenient grouping of compounds. The various generic codes differ primarily in the number and kind of groups they can create, the number of compounds included in each group, and the manner of retrieving data from the coded material.

Rotadex consists of a rotated index with three aspects: molecular formulas, generic structural codes, and the addresses where the references are made to the compounds. Any combination of the structural code and elemental composition may be used to define or narrow down a field of search. Each Rotadex structural code contains the same number of characters. In the examples described below, four-character structural codes will be used. In Fig. 1 we see the chemical structure of N,N-dibutylacrylamide, together with its molecular formula, $C_{11}H_{21}NO$, and its Rotadex structural code 99QT.

$$Bu_2N\text{-}\overset{\overset{\displaystyle O}{||}}{C}\text{-}CH=CH_2$$

$C_{11}H_{21}NO$ 99QT

Fig. 1.—N,N-Dibutylacrylamide

Each character of the structural code is obtained from a separate table. The assignment of chemical features to these tables is tentative and the following should be considered only as examples. The first character of the code 99QT comes from Table I. The presence or absence of any of five chemical features may be designated by one of the 32 alphanumeric characters given in the right-hand column. Note that the letters in the body of the table are merely mnemonic abbreviations for the chemical features. In the table, r represents any homocyclic ring (*e.g.*, cyclopentane), 2r any homocyclic two-ring fusion (*e.g.*, naphthalene), 3r any homocyclic three-ring fusion (*e.g.*, anthracene), 4r any homocyclic four(or greater)-ring fusion (*e.g.*, steroids), and sp any spiro configuration. It is thus seen that a benzene-substituted naphthalene would be described by the letter P taken from this table. Note that 2r does not indicate two separate benzene-like rings, but rather one or more two-ring fusions.

Table I

Chemical features					Code
r					A
r				sp	B
r			≧ 4r	sp	C
r			≧ 4r		D
r		3r	≧ 4r		E
r		3r	≧ 4r	sp	F
r		3r		sp	G
r		3r			H
r	2r	3r			I
r	2r	3r		sp	J
r	2r	3r	≧ 4r	sp	K
r	2r	3r	≧ 4r		L
r	2r		≧ 4r	sp	M
r	2r		≧ 4r	sp	N
r	2r			sp	O
r	2r				P
	2r				Q
	2r			sp	R
	2r		≧ 4r	sp	S
	2r		≧ 4r		T
	2r	3r	≧ 4r		U
	2r	3r	≧ 4r	sp	V
	2r	3r		sp	W
	2r	3r			X
		3r			Y
		3r		sp	Z
		3r	≧ 4r	sp	3
		3r	≧ 4r		4
			≧ 4r		6
			≧ 4r	sp	7
				sp	8
-	--	--	--	--	9

The digit 9 is used here (and in the remaining tables) to signify the absence of all five of the chemical features covered by a character of the structural code. The first 9 in the code for N,N-dibutylacrylamide describes, therefore, the absence of any homocyclic ring structure or spiro configuration.

In Table II we see an analogous table from which the second character of the structural code is derived. The second 9 in the code 99QT indicates the absence of any heterocyclic ring structure or bridge configuration in the compound.

Table III is used for the third character of the structural code which describes five chemical configurations of oxygen or sulfur (with no differentiation between them). This character indicates the presence of any acid group

* Presented before the Division of Chemical Literature, 143rd National ACS Meeting, Cincinnati, Ohio, January 14, 1963.

Table II

Chemical features					Code
h					A
h				br	B
h			≧4h	br	C
h			≧4h		D
h		3h	≧4h		E
h		3h	≧4h	br	F
h		3h		br	G
h		3h			H
h	2h	3h			I
h	2h	3h		br	J
h	2h	3h	≧4h	br	K
h	2h	3h	≧4h		L
h	2h		≧4h		M
h	2h		≧4h	br	N
h	2h			br	O
h	2h				P
	2h				Q
	2h			br	R
	2h		≧4h	br	S
	2h		≧4h		T
	2h	3h	≧4h		U
	2h	3h	≧4h	br	V
	2h	3h		br	W
	2h	3h			X
		3h			Y
		3h		br	Z
		3h	≧4h	br	3
		3h	≧4h		4
			≧4h		6
			≧4h	br	7
				br	8
-	--	--	---	--	9

Table III

O ‖ C—O	O ‖ C	OH | C	| O |	O ‖ =O	Code
ac					A
ac				diox	B
ac			ox	diox	C
ac			ox		D
ac		ol	ox		E
ac		ol	ox	diox	F
ac		ol		diox	G
ac		ol			H
ac	one	ol			I
ac	one	ol		diox	J
ac	one	ol	ox	diox	K
ac	one	ol	ox		L
ac	one		ox		M
ac	one		ox	diox	N
ac	one			diox	O
ac	one				P
	one				Q
	one			diox	R
	one		ox	diox	S
	one		ox		T
	one	ol	ox		U
	one	ol	ox	diox	V
	one	ol		diox	W
	one	ol			X
		ol			Y
		ol		diox	Z
		ol	ox	diox	3
		ol	ox		4
			ox		6
			ox	diox	7
				diox	8
--	---	--	--	----	9

(ac whether free, salt, or ester), carbonyl group (one), hydroxyl group (ol), bond—O—bond (ox), and the presence of two oxygens or sulfurs attached through double bonds to the same element (diox). The Q in the code 99QT thus shows that, of these substructures, only one is present.

The fourth character in the structural code, Table IV, describes nitrogen or phosphorus as being present in tertiary form, az, with at least one hydrogen attached, am, in the N-N or P-P configuration, diaz, or in a quaternary state, +. The chemical feature diaz is also used to indicate any contiguous repeated hetero-elements. Examples of this include: S-S, O-O, B-B, etc. The fourth character is also used to show the presence of any non-resonating unsaturations, ene. These unsaturations are named in any case where they are not explicitly included in another chemical feature, that is to say, ac, one, diox, do not call for renaming the ene, but all other instances do. The T in the code for N,N-dibutylacrylamide therefore indicates a tertiary nitrogen *and* a double or triple bond, az, ene.

The addition of a fifth character to the Rotadex structural codes might be useful if it were desirable to indicate additional chemical features such as the presence of metallic salts, organic salts, polymers, incompletely-known structures, etc. With four characters, one may write 32^4 or over a million different structural codes. Though almost everyone of these codes is theoretically acceptable, they will not, or course, find equal use in actual applications. Further, it is from a study of these frequencies that will come improved assignment of chemical features to the tables.

The Rotadex structural codes were designed to facilitate the unambiguous assignment of compounds to proper categories. There is little double-naming of substructures. When a spiro configuration is present the individual moieties are named independently in addition to the spiro designation. For example, spiropentane (*cf.* Table I) would be indicated by the letter B in the first position of the structural code, specifying the presence of the homocyclic rings, r, as well as the spiro configuration, sp. Similarly, Table II, the component rings of a bridge (after eliminating the bridge) are coded in addition to the bridge, br, itself. In order to make the naming of the component rings of a bridge unambiguous, a hierarchy must be established for the elimination of "arms" from the structure. Homo-element arms (carbon) are eliminated first, always starting from the shortest and moving toward the longest arms. Thereafter hetero-element arms are eliminated, again starting from the shortest arm. This elimination procedure stops as soon as the remaining structure can no longer be termed a bridge compound.

Table IV

$\begin{array}{c}\mathrm{H}\\ \vert\\ \mathrm{N}\end{array}$	$\equiv$	+	N	$\begin{array}{c}\mathrm{N}\\ \vert\\ \mathrm{N}\end{array}$	Code
am					A
am				diaz	B
am			az	diaz	C
am			az		D
am		+	az		E
am		+	az	diaz	F
am		+		diaz	G
am		+			H
am	ene	+			I
am	ene	+		diaz	J
am	ene	+	az	diaz	K
am	ene	+	az		L
am	ene		az		M
am	ene		az	diaz	N
am	ene			diaz	O
am	ene				P
	ene				Q
	ene			diaz	R
	ene		az	diaz	S
	ene		az		T
	ene	+	az		U
	ene	+	az	diaz	V
	ene	+		diaz	W
	ene	+			X
		+			Y
		+		diaz	Z
		+	az	diaz	3
		+	az		4
			az		6
			az	diaz	7
				diaz	8
--	---	-	--	----	9

At this point the remaining ring configurations are coded. The bridge is designated whether the compound is homo- or heterocyclic.

Larger chemical features are coded rather than their component parts. For example, acid groups are not also coded for the one or ol groups.

Acid groups and diox configurations are considered the largest of such features since they involve at least three groups of atoms.

In the case of ties, priority is assigned to chemical features in the order in which they appear in the coding tables. Thus, the sulfo radical is coded as ac, one rather than diox, ol.

The ox feature, Table III, includes ethers and epoxy compounds. The epoxy structure is double-coded for a heterocyclic single ring, h, as well as for the ox. Epoxy structures constitute the only case where hetero-elements contained completely within a ring are double-indicated. Otherwise hetero-element substructures are indicated only when all or part of the substructure lies outside the ring. For example, large cyclic ethers are named only as a heterocyclic ring, but cyclic ketones have the one indicated since this extends outside the ring. Similarly, the am in piperidine *is* indicated (since the hydrogen extends outside the ring) and so for diaz when one of these nitrogens is outside the ring.

Hydrazine derivatives are always coded as diaz and, whenever one or more hydrogens remain unsubstituted, the derivatives are double-coded am also.

Since the quaternary charge implies the presence elsewhere in the compound of a negative charge, all quaternaries, phosphoniums, sulfoniums, etc., are indicated whether appearing in a ring or not.

It should be noted that because of the ancillary use of the molecular formula, the Rotadex structural code may conveniently group together chemical features which are coded differently by other codes.

Rotadex does not distinguish between substitutions located on rings or located on chains, nor does it signify the exact number of times a chemical feature appears in a compound.

Preliminary tests suggest that a four-character structural code may be assigned to the average compound in less than half a minute. This attribute and the terseness of the code make it particularly suitable for incorporation in large-scale coding and printing operations.

Generic searching with Rotadex is facilitated by a fixed-column format.[24] This assures the user that a scan of a specific column will disclose all references to a given feature, whether it be part of the elemental composition or the structural code.

"Rotation" of molecular formulas[25a] and chemical codes are performed without actually moving the characters out of their fixed-column positions. Rather, the formulas are sorted successively on each element and the codes on each character. Whichever position constitutes the primary sort key, the elements or characters to the right constitute successive, minor sort keys. Minor sort keys to the right of the last character continue with the first character (in a wrap-around fashion). A primary sort on hydrogen is omitted (Fig. 2). The order of the elements in the fixed columns reserved for molecular formulas is: C, H, N, O, P, S, X (all halogens), and M (all remaining miscellaneous elements).

C	H	N	O	P	S	X	M	Code	Address
6	7	2	4		2	F 1		A98A	3743-5
7	9	2	4		2	F 1		A98A	3743-6
8	11	2	4		2	F 1		A98A	3779-4
								A98A	3779-50
9	13	2	4		2	F 1		A98A	3779-9
8	5	2	4		2	F 3		A976	8503-4
9	5	2	4		2	F 5		A976	8503-6
12	28	2	4		2	F 6	Si	99AH	1557-13
10	5	2	4		2	F 7		A976	8503-9
11	6	2	4		2	F 8		A976	8503-16
11	4	2	4		2	F10		A976	8503-21
13	4	2	4		2	F14		A976	8503-23

Figure 2.

Within the X and M columns, the various elements are alphabetized. Since common elements appear in fixed columns (C through S), only the number of atoms of each of these elements need be listed, and letters may be replaced by headers of the atomic symbols over the columns. Allowances are made for maximal, usual number of digits required for these elements. Thus, two column positions are reserved to indicate the number of carbons, two for the number of hydrogens, and two for oxygen. One column each is reserved for the elements nitrogen, phosphorus, and sulfur. The halogen and miscellaneous fields each include three columns since they must also contain up to two alphabetical characters indicating the elements. In any case, when the actual number of digits exceeds the allotted columns, the excess lower-order digits are dropped to the next lower line where they and a lozenge are printed in an otherwise blank line.

The print-out of this molecular formula aspect of the index reveals which column is the primary sort key by the obvious grouping in that area. For instance, when the sort key has moved to nitrogen, there appears a listing of all the molecular formulas headed by a block of blanks in the nitrogen column followed by a continuous string of 1's followed by a continuous string of 2's, etc. Figure 2 illustrates what a fragment of the actual index may look like and is taken from a hypothetical section where nitrogen is the primary sort key.

Searches are best performed with Rotadex by first defining the minimal, elemental requirements and parameters of acceptable structural, codes. In some searches, the presence of a relatively rare element provides a convenient starting point. Any search involving boron, for instance, would probably best begin at the portion of the index where the data has been sorted first on the miscellaneous elements. There, under the B's, the user will find all compounds containing boron and within this group the compounds will be arranged by the number of borons present. From this point, the user may scan other elemental requirements and verify an acceptable structural code before looking up specific compounds at their addresses.

Many searches will allow the user to subdivide the total file by scanning elements successively from left to right. Final selection of appropriate compounds can again be made by the acceptable structural codes.

Whenever cumulative molecular formula indexes are compiled, some molecular formulas will accrue unwieldy numbers of compound addresses. These clusters are discouraging to the user since many or most of the compounds in a cluster differ from the type of compound desired. The inclusion of the Rotadex structural code with the address of each compound allows for further subdivision of these clusters in accordance with the chemical features required by the search. Figure 3 shows a sampling of the compounds from the 36 addresses clustered under the common molecular formula $C_{11}H_{21}NO$, as it appeared in a biennial cumulative index of *Index Chemicus.* When Rotadex structural codes are assigned to these compounds, two cases of identical codes are encountered. In the first instance, a series of stereoisomers of decahydroquinolines all receive the same structural code 9QY9, and the second case turns out to be separate references to the identical compound, N,N-dibutylacrylamide. The remaining compounds receive differing structural codes.

Another method of searching Rotadex calls for first entering the portion of the index where compounds have been sorted primarily on the structural codes. Table V is a summary of the four preceding tables, and it is used to decide which codes will be compatible with a given search. We see that any given chemical feature may have been coded (in the appropriate column) by any of 16 characters, depending upon what other chemical features are also found within the compound.

If a generic search were to be conducted for all N,N disubstituted acrylamides, the minimal elemental requirements of C_3H_3NO would be of little help. The structural code, however, would narrow the search considerably. No restrictions would be imposed on the first or second characters of the structural code in this case, but the third character must be one of the letters of the alphabet I–X since there must be at least one =O present in any acrylamide. Note that the Q in the code 99QT meets this requirement. Furthermore, the fourth character must indicate a double bond *and* a tertiary nitrogen. The only characters which meet the requirement for a

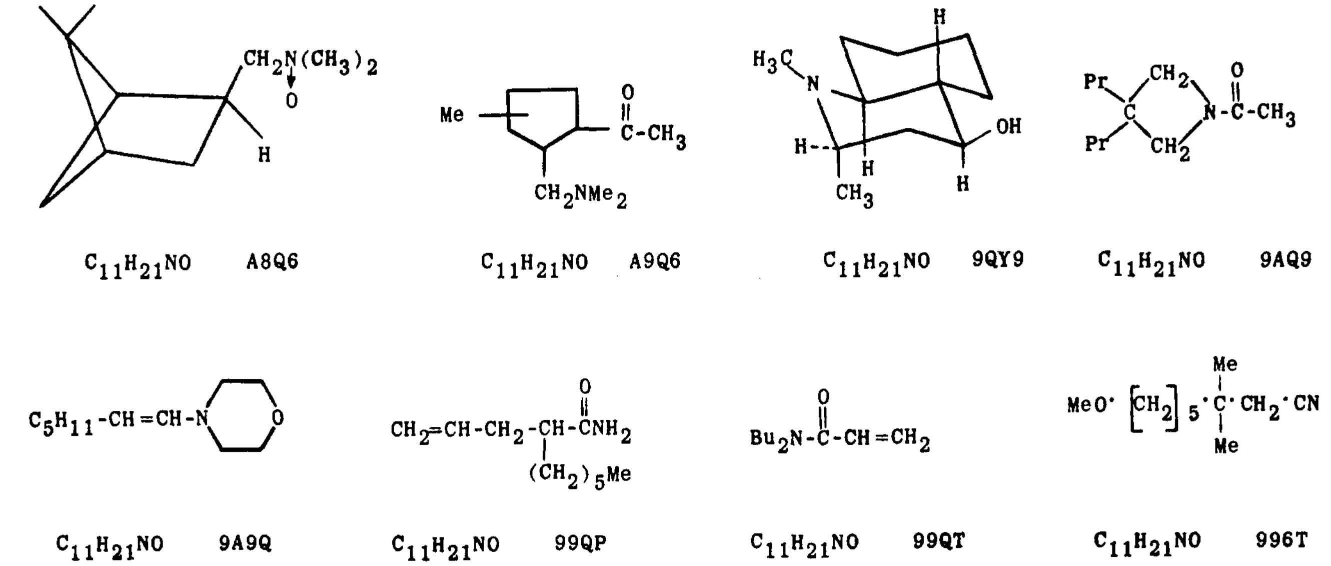

Figure 3.

Table V

Character #	1	2	3	4	
Chemical features	Homo-cyclics, spiros	Hetero-cyclics, bridges	O and S	N and P ≡	Structural code characters
	r	h	ac	am	A–P
	2r	2h	one	ene	I–X
	3r	3h	ol	+	E–L, U–4
	≧ 4r	≧ 4h	ox	az	C–F, K–N, S–V, 3–7
	sp	br	diox	diaz	BC, FG, JK, NO, RS, VW, Z3, 78
	--	--	----	----	9

double bond (I–X) together with tertiary nitrogen (C–F, K–N, S–V, 3–7) are K–N, S–V. Note that the T of the code 99QT meets this requirement.

The generic search for N,N disubstituted acrylamides might, therefore, be performed by first entering the index at the section where compounds have been sorted on the fourth character of the structural code. All entries in the two alphabetical portions, K–N and S–V, would then be scanned for a character in the third position equal to I–X. All N,N disubstituted acrylamides would be included in these results. The amount. of "noise" or extraneous compounds also included, will vary with individual searches.

In summary, Rotadex is a new proposal for indexing and listing chemical compounds. The data is sorted repeatedly on each part of the molecular formulas and structural codes. When printed in fixed-column format, this enables the user to narrow the range of his search rapidly according to many combinations of restricting parameters.

The structural codes devised for Rotadex are terse and simple. The chemical features described by the structural code may be combined as desired to form relatively broad descriptors of the chemical structures. Clusters appearing under identical molecular formulas are subdivided by the structural codes.

Structural codes may be entered directly to perform generic searches.

Though designed to facilitate hand searches of printed indexes, Rotadex is also well-suited to mechanical and computer searches.

Coding and searching of compounds is independent of systems of nomenclature and can be performed very rapidly. Only a few minor hierarchical rules are required. Chemical features to be coded are mutually exclusive. Substructures may be described by the fragmentary chemical features which they contain.

It should also be possible to write a computer program that will automatically convert systematic chemical names of compounds into corresponding molecular formulas and structural codes by an extension of Garfield's[25b] algorithm for translating chemical nomenclature.

REFERENCES

(1) "A Method of Coding Chemicals for Correlations and Classification," National Research Council, Washington, D. C., December 12, 1949.

(2) H. T. Bonnett and D. W. Calhoun, *J. Chem. Doc.*, **2**, 2 (1962).

(3) A. Cahn, Abstracts, 142nd National Meeting of the American Chemical Society, Atlantic City, N. J., Sept., 1962, p. 4G.

(4) E. M. Crane and M. M. Berry, *Chem. Eng. News*, **33**, 2842 (1955).

(5) E. Dale and K. Heumann, "Statistical Information on Component Parts of Chemical Compounds," Chemical-Biological Coordination Center, National Academy of Sciences-National Research Council, March, 1955.

(6) W. M. Duffin, *J. Chem. Doc.*, **1** (3), 44 (1961).

(7) G. M. Dyson and E. F. Riley, *ibid.*, **2**, 19 (1962).

(8) A. Feldman, D. B. Holland, and D. P. Jacobus, *ibid.*, **3**, 187 (1963).

(9) J. Frome and J. Leibowitz, "A Manual for Coding Steroids," Patent Office Research and Development Report No. 11, November 17, 1958.

(10) J. Frome, *et al.*, "Manual for a Punched Card Retrieval System for Organic Phosphorus Compounds," Patent Office Research and Development Report No. 22, November 24, 1961.

(11) J. Frome and P. T. O'Day, *J. Chem. Doc.*, **2**, 248 (1962).

(12) J. Frome, *et al.*, "ASTIA Chemical Thesaurus," Armed Services Technical Information Agency, Arlington, Va., December, 1962.

(13) A. Gelberg, W. Nelson, G. S. Yee, and E. A. Metcalf, *J. Chem. Doc.*, **2**, 7 (1962).

(14) M. Gordon, *et al.*, "Chemical Ciphering," Royal Institute of Chemistry of Great Britain and Ireland, Proc. XIth Intern. Congr. Pure Appl. Chem., London (1947), Vol. II, Sec. III, p. 115.

(15) H. W. Hayward, "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," Patent Office Research and Development Report No. 21, 1961.

(16) T. R. Norton, "A Manual for Coding Organic Compounds," unpublished paper, May 27, 1953.

(17) J. O'Connor, "A Note on the Possibility of a Divided Structure File Permitting Arbitrary Substructure Searches," University of Pennsylvania, Philadelphia, Pa.

(18) "Rules for IUPAC Notation for Organic Compounds," John Wiley & Sons, Inc., New York, N. Y., 1961.

(19) J. Silk, *J. Chem. Doc.*, **3**, 189 (1963).

(20) E. G. Smith, *Science*, 131 (1960).

(21) P. E. Verkade, *Chem. Weekblad.*, **58**, 137 (1962).

(22) K. W. Wheller, *et al.*, *Am. Doc.*, **9**, 198 (1958).

(23) W. J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell Co., New York, N. Y., 1954.

(24) J. O'Connor, *Am. Doc.*, **13**, 204 (1962).

(25) E. Garfield, Abstracts, 141st National Meeting of the American Chemical Society, Washington, D. C., March, 1962: (a) p. 8G; (b) p. 7G.