CHEMICAL CODING FOR INFORMATION RETRIEVAL

BY W. M. DUFFIN
The Wellcome Research Laboratories, Langley Court, Beckenham, Kent, England

Chemical coding systems nowadays generally are designed for use with punched cards by either manual or mechanical operation. Large numbers (say, more than ten thousand) of edge-punched cards are unsatisfactory. Mechanically sorted cards have been brought to a high efficiency but are not convenient for use away from the machines. Moreover, it is not normally possible to add detailed information to the cards (a) for want of space and (b) because of the difficulty of extracting each separate card as required.

Our objectives were to index all compounds which had been tested by us, all compounds marketed or patented in the pharmaceutical and veterinary fields, and all chemicals which are commercially available. We also wanted to be able to add to the information indexed so that any card showed on inspection all that was known about the compound. The system here outlined is now dealing satisfactorily with some 70,000 compounds. 5,000 compounds are added each year and some 200 entries per day are made on existing cards. Moreover, by grouping like compounds together, search for one compound automatically reveals others of the same type, and all tests done on them.

The index will not permit the immediate location of a ring system which is not the senior ring present, but experience has shown that this is rarely required, nearly all the enquiries being directed to the senior system. Enquiries directed merely to say ether-groups do not make much sense without some indication of the other groups present, because there are so many of them. An enquiry for all the thiosemicarbazones was answered in three hours, the answer giving not merely their reference numbers (and there were over a hundred), but all the test results. Again such enquiries are rare, this being the only such occasion in three years. Apart from such routine enquiries as to whether a compound has been examined (about ten a day), an average of three a day are for more detailed information. Nearly all are answered by telephone. The index is run by one graduate and one non-graduate with research experience, with such clerical assistance as is needed for typing reports.

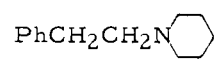
The compounds are filed on cards, or in loose-leaf binders, in alphabetical order of their coding. Loose-leaf binders are preferred, as cards tend to be misplaced or "borrowed."

Each chemical grouping or ring system is allocated two letters and the coding for a particular compound is obtained by assembling these in reverse alphabetical order. In the majority of cases the coding for a group is abbreviated to the first letter, and the second is added only in special cases, the second letter being lower case, e.g., Jq.

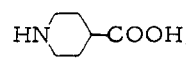
The coding for a compound consists of 3 parts, $\alpha/\beta/\gamma$

α indicates senior ring present (Table I)
β indicates other rings and groups (Table II)
γ indicates total number of carbons present

In order to bring like compounds together, piperidine, morpholine and pyrrolidine (unsubstituted or substituted by alkyl or halogen) in the presence of other ring systems and acting only as tertiary amines theoretically replaceable by $-NMe_2$ without change of chemical type, are not coded by rings, but as amines, as in



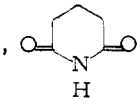
. In the absence of other ring systems, or where any other substituent is attached to a carbon of the ring, e.g.,



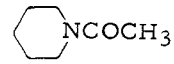
this rule does not apply, and they qualify to be coded as rings.

For a similar reason, methylenedioxy compounds are coded as diethers and not as ring systems.

Each group attached to carbon of a ring is coded individually without reference to other parts of the molecule, e.g.,



is coded as piperidine + 2 carbonyl groups, not as amide. In those cases where an exocyclic group is attached to the hetero-atom of a ring, the hetero-atom is considered as part of the group, e.g.,



is coded as an amide,

as an urea, etc.

In those cases where a group may be considered in isomeric forms, the senior coding is used, e.g., uracil, where the oxygen-containing groups are coded as carbonyl, not hydroxyl.

If the structure is unknown, the coding is X/ followed by the name.

TABLE I

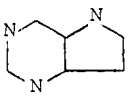
α, Ring System.—In order of seniority [no distinction made between saturated or unsaturated rings, except Th Cyclohexyl, Ve Piperazine]

Wy	Not otherwise included
Ww	ON rings not otherwise included
Wv	OS rings not otherwise included
Wu	NS rings not otherwise included
Wt	S rings not otherwise included
Ws	O rings not otherwise included
Wr	N rings not otherwise included

These letters are followed by two numbers. The first shows the total number of heterocyclic atoms, the second the total number of rings, e.g., phenothiazine, Wu23 pyrrocoline, Wr12

Wq	Tetrazine	Vu	1,2,4-Triazole	Tm	Homocyclic >7C atoms
Wp	Triazine	Vt	1,2,3-Triazole	Tl	Cycloheptane
Wn	Purine	Vr	Pyrazole	Tk	Cyclopropane
Wm	Pyrazine	Vq	Imidazole	Tj	Cyclobutane
Wl	Pyrimidine	Vp	Thiazole	Ti	Cyclopentane
Wk	Pyridazine	Vn	Oxazole	Th	Cyclohexane and Hydrobenzenes
Wg	Acridine	Vh	Carbazole	Tg	Tri- and Tetra-phenylmethane
Wf	Phenanthridine	Vg	Indole	Tf	Diphenylmethane
We	Isoquinoline	Vf	Pyrrole	T_n	n-Unfused benzene rings
Wd	Quinoline	Ve	Piperazine	T	Benzene
Wc	Pyridine	Vd	Thiophen	O	No ring present
Wb	Thiopyran	Vb	Furan		
Wa	Pyran	T^n	n-Fused carbon rings		
Vv	Tetrazole				

Referring to Table I: when benzene rings are the senior ring present, and there are 'n' of them, use the symbol T_n in the α position of the code. In all other cases of duplication of the senior ring, use the simple symbol in the α-position, and indicate the other(s) in the β-position. If it is required to break down groups Wr to Ww further, this may be done either by indicating the "component" nuclei by their separated letters or by indicating the number of heterocyclic atoms in each, e.g.,



Wr32, may be sub-divided as Wl, Vf or as 2,1.

TABLE II

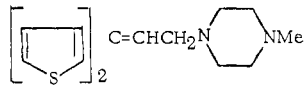
β, Other Rings and Groups.—In order of seniority. Normally only the first letter is used.

W As Table I.
V As Table I.
Ux Element other than C, H, N, O, S or halogen, where x is the symbol of the element. Alkali or alkaline earth metals occurring as salts of acids are not included.
T As Table I. Normally abbreviated to T, but for Tf put T_2, for Tg put T_3 or T_4, for T^n put T^n.
S Sa Thiocyanates; Se isothiocyanates; Sg thiosemicarbazides and thiosemicarbazones; Sj thiourea; Sk - NHCSSH; Sn sulfamic acids.
R Ra Sulfonic acids; Rb sulfonyl halides; Rc sulfonamides; Rd Sulfonates; Rj sulfones; Rm sulfoxides; Rp -COSH; Rq -CSOH; Rr -CSSH; Ru sulfhinic acids; Rx sulfenic acids.
Q Qa Mercaptans; Qc sulfur chlorides; Qe sulfides; Qj disulfides; Qp thioketones.
P Pa Ureas; Pe guanidines; Pf semicarbazides; Pg semicarbazones; Pj urethans; Pp amidines; Pq amidoximes.
N Nd Cyanates; Ne isocyanates; Nf C-nitroso; Ng N-nitroso; Nj hydroxylamines or N-oxides; Nk oximes; Np hydroxamic acids.
M Ma Nitro; Mz nitramine.
L La Cyanides; Lm isocyanides; Ls cyanamides.
K Ka Diazonium compounds; Kb azo compounds; Kc diazoamino; Kd azoxy; Kh hydrazines and hydrazides; Km acid azides; Kn diazoketones; Kp azides.
J Ja Amines; Jl N-haloamines; Jq amides; Jv proteins and peptides; Jy imino-ethers; Jz imino-chlorides.
H Ha Esters of carboxylic acids; He acid halides; Hj acid anhydrides.
G Ga Carboxylic acids.
F Fa Oxides; Fb ozonides; Fc peroxides; Fe ethers; Ff acetals; Fm ketenes; Fs carbohydrates; Ft glycosides.
E Ea Carbonyl; Ez quinones.
D Da Hydroxyl.
C Ca Halides.
B Ba Olefinic bonds in chain; Bc acetylenic bonds in chain.
A Alkyl group (with no codable substituent) attached to C of ring. If a grouping occurs more than once, a subscript number is added to the symbol (see Examples).
"Onium" compounds are indicated by a superscript ^{4}, e.g., J^4 ammonium; C^4 iodonium.

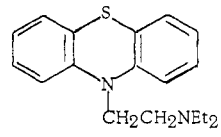
EXAMPLES

CH_3CH_2OH O/D/2

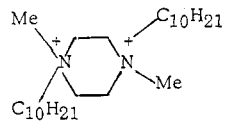
PhCH(OH)CHMeNHMe T/JD/10



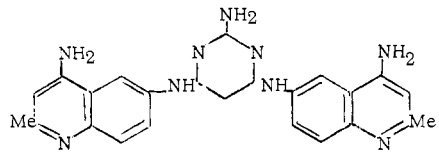
Ve/V_2B/16



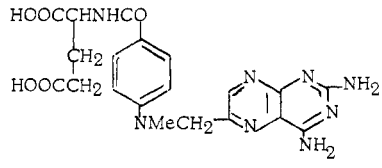
Wu23/J/18



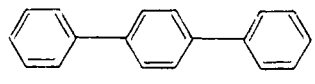
Ve/J_2^4/26



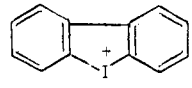
Wl/$W_2J_5A_2$/24



Wr42/TjqJ$_3$G$_2$/20



T_3/18



Wy13/C^4/12

It has been found convenient to treat phosphorus compounds and others of the U class in an arbitrary rather than strictly chemical manner. Phosphorus esters P(OR) are coded F (as if they were ethers), PS and P(SR) as Q and $P(NH_2)$ as J. P(O), P-O-P and PCl are ignored.

Example: $(EtO)_2P(S)O[CH_2]_2SEt$ O/UpQ$_2$F$_3$/8

In use, if details of a particular compound are required, the full coding is worked out and the record is found immediately. Homologs, differing only in the number of carbon atoms, will be found on each side of it. If more general information is required, or it is required to code for a wide group of compounds (e.g., in patents), the coding is taken only far enough to cover the general class. For example, phenylpyrimidines are covered by W1/T and under this heading will be found general references covering this series of compounds. Following it, in order of coding, will be found references where more detailed coding is possible, e.g., W1/TJ for amino derivatives, and finally the individual compounds, e.g., pyrimethamine at $W1/TJ_2CA/12$.

Because of the ease with which a particular compound can be located (seconds only) it is a simple matter to add further information to an existing sheet, or to see what analogs have been examined. A mechanically sorted punched card installation which has been maintained for 14 years is so much slower than the master index that it is to be abandoned.

The cards used were standard I.C.T. 65-column cards, with a column allotted to each item of the β-coding, and a group of columns provided for the α-coding. This system was of use for location of "junior" groupings but took a considerable time because all or nearly all the cards had to be searched, and when the search was done the numbers obtained had to be referred to the manual index to get the test results. The chief use made of the punched cards in recent years has been the production of a complete index each year (in order of coding) which was distributed to the various research laboratories for their information. The cessation of this index will be a considerable loss, but is being overcome by more efficient arrangements for answering questions from the manual index by telephone. Again, experience has shown that many people have preferred to ask the central index--it is less trouble to themselves, it is up-to-date, and it contains all the information.

This coding also is being used for a classified index to reactions. For this, six letters are used. For example, GaJq/Ja includes all those methods by which a carboxylic acid (Ga) is converted into an amide (J_q) by reaction with an amine (Ja).

For this purpose A and Z letters as "operators" are used (Table III). Y letters are used for biological activities.

TABLE III

Aa	Reduction	Am	Nitration	Za	Characterization
Ab	Dehalogenation	An	Nitrosation	Zb	Stability
Ac	Halogenation	Ap	Disproportion	Zc	Optical Resolution
Ad	Hydrolysis	Aq	Sulfur	Ze	Physical Properties
Ae	CO	Ar	Sulfonation	Zm	Metabolism
Af	Oxidation	As	Desulfurization	Zn	Nomenclature
Ag	CO_2	Au	Metals	Zs	Structure
Ai	Isomerization	Aw	Dehydration	Zx	Isolation
Aj	Ammonia	Ax	Polymerization	Zy	Availability
Al	HCN or $(CN)_2$	Az	Degradation	Zz	Methods of synthesis

SYNTACTIC STUDY OF CHINESE AND ENGLISH

A comparative syntactic study of the Chinese and English languages will be undertaken with a NSF grant to the Ohio State University Research Foundation. The purpose of the study is to facilitate machine translation and information retrieval.

BRIDGMAN RESEARCH PAPERS

The collected research papers of the late Dr. Percy W. Bridgman will be published by Harvard University with the assistance of the National Science Foundation. The collection consists of about 200 papers to be published in seven volumes.