Table I. U. S. Production of Organic Chemicals, Millions of Pounds (1965)

Chemical	Year Forecast Published	Estimate and Reference	U.S.T.C. Figures	Difference	Accuracy of Estimate, %
Acetic acid	1965	1100[a]	1346	−246	81.7
Acetic anhydride	1965	1430[a]	1531	−101	93.4
Acetone	1965	1135[a]	1124	+11	99.0
Acrylonitrile	1962, 1965	450[b], 670[a]	771.6	−321.6, −101.6	58.3, 86.8
Aniline	1965	170[a]	195.5	−25.5	86.4
Benzoic acid	1962	12[c]	16.2	−4.2	74.1
Carbon disulfide	1965	690[a]	756.5	−66.5	91.3
Carbon tetrachloride	1965	570[a]	593.6	−23.6	96.0
Cyclohexane	1963	1200[d]	1700	−500	70.6
Decyl alcohol	1961	75[e]	105.9	−30.9	70.8
Ethylene glycol	1965	2000[a]	1798	+202	89.9
2-Ethylhexanol	1961	210[e]	293.2	−83.2	71.6
Formaldehyde	1965	2770[a]	3106	−336	89.2
Iso-octyl alcohol	1961	85[e]	126.7	−41.7	67.1
Isopropyl alcohol	1963, 1964	1400[f], 1640[g]	1540	−140, + 100	90.9, 93.9
Monosodium glutamate	1961	30[h]	43.1	−13.1	69.6
Pentaerythritol	1965	72[a]	69.3	+2.7	95.8
Phthalic anhydride	1964, 1965	550[i], 650[j]	608.3	−58.3, +41.7	90.4, 93.5
Propylene glycol	1965	265[a]	212.8	+47.2	80.3
Propylene oxide	1963	480[f]	604.6	−124.6	79.4

[a] *Chem. Eng. News* **43** (1), 12, 1965. [b] *Ibid.* **40** (53), 11, 1962. [c] *Ibid.* **40** (46), 34, 1962. [d] *Oil Gas J.* **61** (6), 107, 1963. [e] *Chem. Eng. News* **39** (46), 129, 1961. [f] *Oil Gas J.* **61** (23), 202, 1963. [g] *Chem. Eng. News* **42** (49), 26, 1964. [h] *Ibid.* **39** (49), 29, 1961. [i] *Chem. Week* **94** (3), 59, 1964. [j] *Chem. Week* **96** (2), 79, 1965.

Table II. Accuracy of Estimates by Year of Publication

	Underestimates, Degree of Accuracy			Overestimates, Degree of Accuracy		
Year	<70%	70-80%	>80%	80-90%	90-100%	Total Estimates
1965	...	...	7	2	3	12
1964	...	...	1	...	1	2
1963	...	2	1	...	...	3
1962	1	1	...	...	...	2
1961	2	2	...	...	...	4
Total Estimates	3	5	9	2	4	23

at least 80% accurate. The six predictions made in 1961 and 1962 were less than 80% of the government figures.

Of the 18 estimates from *Chemical & Engineering News*, 12 were more than 80% in agreement with the Tariff Commission figures. Six of these forecasts were at least 90% of the government statistics.

LITERATURE CITED

(1) Pafford, P. T., Division of Chemical Marketing, 157th Meeting, ACS, Minneapolis, Minn., April 1969.

The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems

J. H. R. BRAGG,* M. F. LYNCH, and W. G. TOWN**
Postgraduate School of Librarianship and Information Science, University
of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received December 4, 1969

The development of large computer files of chemical structures has increased the need for efficient registration techniques. In a recent paper, Lynch *et al.*[1] have described an extension to the isomer sort technique in which registration is accomplished without generating a unique description for the compound being registered.

The technique depends on prediction of the approximate size of the group of chemical compounds having the same molecular formula as the compound being registered in a particular file. This is then used to decide the appropriate level of description for the compound; in general, the larger the molecular formula group, the deeper is the level of description necessary to obtain subgroups of optimum size. In this paper, we describe a method of predicting molecular formula group sizes from the molecular formula itself and the known file characteristics.

*Present address: European Research Centre, Texaco (Belgium), Ghent, Belgium.

**Present address: University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

J. H. R. BRAGG, M. F. LYNCH, AND W. G. TOWN

An analysis of the distribution of chemical compounds among molecular formula groups has been made for the Sixth Collective Formula Index to *Chemical Abstracts*. The larger molecular formula groups exhibit a regular pattern, and simple rules may be devised to obtain estimates of the number of compounds having the same molecular formula in a file of given size and composition. These estimates of molecular formula group size may be used to decide the level of description necessary for a compound during registration by a recent extension of the isomer sort technique.

MOLECULAR FORMULA DISTRIBUTION

A necessary preliminary to the prediction of molecular formula group sizes was an analysis of the distribution of chemical structures among molecular formula groups. We studied the Sixth Collective Formula Index to *Chemical Abstracts* which includes all compounds indexed during the years 1957–61. This was the most convenient source of information for our analysis, as it contains the chemical structures organized into molecular formula groups. This collection is a general one, and the distribution may be different in more specialized collections. However, it is desirable to repeat the analysis, in any case, before the technique is applied to a particular collection of structures and, from time to time, as the collection grows (a simple matter if the collection is already in machine-readable form).

The analysis of molecular formula group distribution was performed by counting the names of the compounds in each group. This tended to produce an underestimate of the populations of the groups, as references to positional isomers are collected under one name in this index. For example, the molecular formula group $C_{10}H_{16}O_2$ contains 251 different names, which represent 329 different structures. The larger the molecular formula group the greater will be the underestimate of the population.

As we were chiefly interested in finding large groups, we first obtained a rough estimate of the population distribution for groups containing more than ten compounds. Population counts were made on all molecular formula groups containing ten or more names in a 4% sample of the pages in the index. In addition, the total number of names on each sample page was obtained. Estimates were then made of the frequency of occurrence of each group size and the total populations in each size of group. The sample pages contain approximately 20,000 names, and hence the whole formula index is estimated to contain some 500,000 names. However, these compound names may represent many more actual compounds.

From the population distribution, shown in Figure 1, it was estimated that approximately 43% of the compounds in the file occur in groups containing more than ten compounds and that 32% are found in groups of 20 or more compounds. The average group size was not determined, but from the part of the distribution studied it would appear to be less than 11 compounds. This may be compared with an average group size of 17 compounds found by Bernays[2] for a file containing 1,500,000 compounds.

The object of the main part of the analysis was to determine the patterns of occurrence of the large molecular formula groups. To reduce the labor involved we concentrated on groups containing 20 or more compounds. The frequency distribution of these groups at different carbon numbers is similar to the distribution of compounds found in the CAS Registry File by Leighner and Leiter[3]. Both distributions show a peak at a carbon count of 14. Therefore, only molecular formula groups with carbon counts between C_{10} and C_{16} were considered. However, the molecular formula groups $C_xH_yO_z$ (where $z = 1 \rightarrow 6$) were considered at all carbon numbers.

The populations of the groups with fixed combinations of hetero-atoms (i.e., atoms other than carbon or hydrogen) were plotted in tabular form, the two axes corresponding to varying numbers of carbon and hydrogen atoms. Examples of these tables are shown in Figures 2 and 3 for the molecular formula groups C_xH_yNO, and C_xH_yO. Blank spaces in the table indicate less than 20 compounds. The heavy line shows the maximum degree of saturation normally possible in a compound, entries above this line being zero.

In many of the tables the population distributions conform to a similar pattern. The greatest populations occur in a triangular region defined by C_xH_y where $x \leqq 22$ and $8 \leqq y \leqq$ saturated hydrogen count. At higher carbon numbers the large groups occur in a band, the center of which is parallel to the line of total saturation. The band is displaced from this line by an amount which depends on the number and types of hetero atoms present. When the band is extrapolated back into the triangular region below C_{22}, it is often found that the most highly populated groups lie within the band.

PREDICTION OF MOLECULAR FORMULA GROUP SIZE

The patterns of population distribution we have obtained can be used to predict the approximate size of a molecular formula group. In the collection studied,

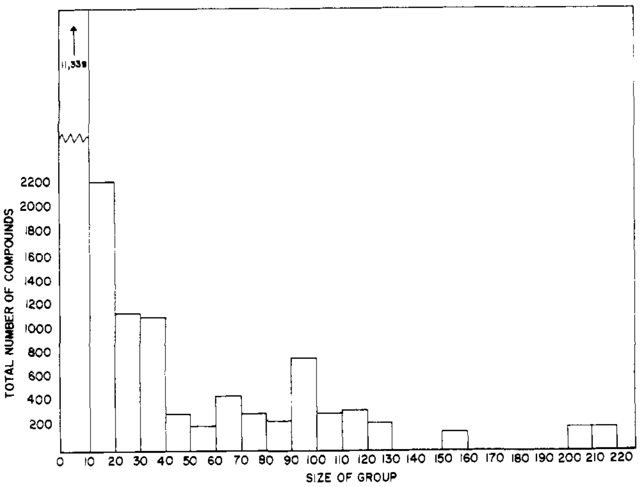
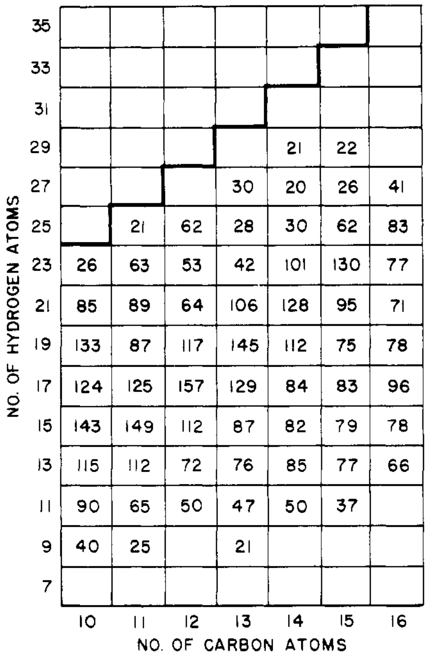


Figure 1. Molecular formula group size distribution in the 4% sample



	10	11	12	13	14	15	16
35							
33							
31							
29					21	22	
27				30	20	26	41
25		21	62	28	30	62	83
23	26	63	53	42	101	130	77
21	85	89	64	106	128	95	71
19	133	87	117	145	112	75	78
17	124	125	157	129	84	83	96
15	143	149	112	87	82	79	78
13	115	112	72	76	85	77	66
11	90	65	50	47	50	37	
9	40	25		21			
7							

NO. OF HYDROGEN ATOMS

NO. OF CARBON ATOMS

Figure 2. Population (≧20) of molecular formula groups in the series C_xH_yNO.

the types and combinations of hetero atoms present in the large molecular formula groups are limited and consist mainly of oxygen and nitrogen, alone or in combination. Within the groups of molecular formulas with a particular combination of hetero atoms, the larger groups occur in certain well-defined regions. The population tables may be regarded as three dimensional surfaces, and the problem reduces to finding mathematical expressions for the contours which delimit the group sizes at which the various levels of description are necessary. The actual limits will be determined by the total file size and the characteristics of the computer system itself, but the technique is not very sensitive to these limits. It will often be sufficient to define the limits of the regions in terms of straight line relationships. For example, the region defined by $x \leqq 16$ and $y \geqq 14$ in Figure 3 contains most molecular formula groups with more than 50 compounds. Although this will give only an approximate definition of the sizes of the molecular formula groups, this will be adequate in most cases. A balance must always be made between the ease of computing the limits of each region and the range of the group sizes contained within it.

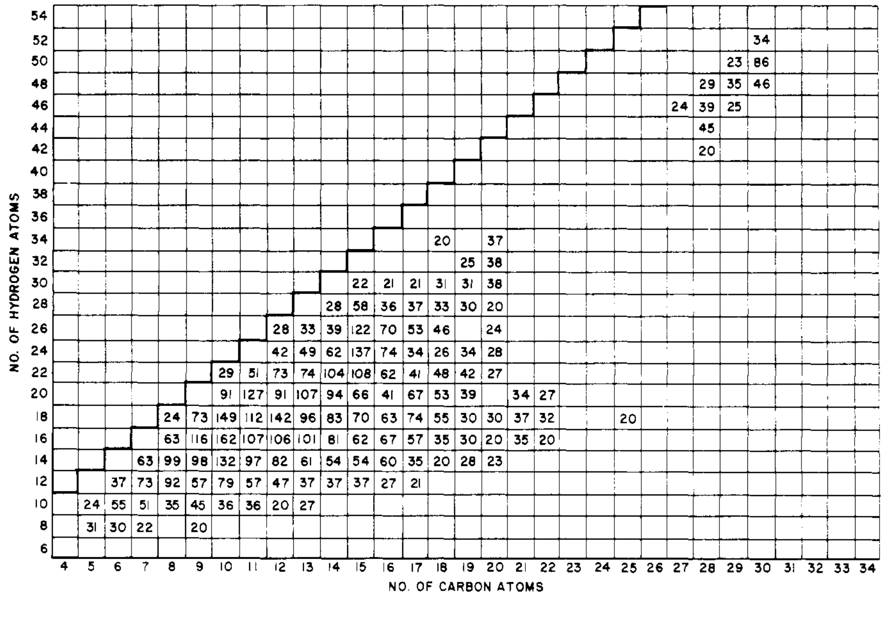
Thus we have shown that, given the total number of compounds in the file and the molecular formula distribution, it is possible to predict the range of group sizes to be expected from the molecular formula alone.

USE OF MOLECULAR FORMULA GROUP SIZE IN REGISTRATION

In the registration technique described by Lynch *et al.*,[1] the first stage in the process is to determine the approximate number of compounds in the file with the same molecular formula as the compound being registered. This is used to decide the tactics to be adopted during the remainder of the registration process. When the molecular formula group is small, it is practicable to determine identity by comparison of the candidate compound with each of the other members of the group. In larger groups, however, it is necessary first to choose the level of description which will divide the molecular formula group into subgroups of optimum size. The subgroup to which the compound belongs is then found and comparison with its members is made. Thus it is possible to determine rapidly the presence or otherwise of the compound in the file, a characteristic which is desirable in registry systems.

CONCLUSIONS

We have analyzed the distribution of compounds among molecular formula groups in a large collection of chemical structures and have demonstrated that molecular formula



	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
54																															
52																											34				
50																										23	86				
48																									29	35	46				
46																								24	39	25					
44																									45						
42																									20						
40																															
38																															
36																															
34															20		37														
32																25	38														
30												22	21	21	31	31	38														
28											28	58	36	37	33	30	20														
26									28	33	39	122	70	53	46		24														
24									42	49	62	137	74	34	26	34	28														
22							29	51	73	74	104	108	62	41	48	42	27														
20							91	127	91	107	94	66	41	67	53	39		34	27												
18					24	73	149	112	142	96	83	70	63	74	55	30	30	37	32			20									
16					63	116	162	107	106	101	81	62	67	57	35	30	20	35	20												
14				63	99	98	132	97	82	61	54	54	60	35	20	28	23														
12			37	73	92	57	79	57	47	37	37	37	27	21																	
10		24	55	51	35	45	36	36	20	27																					
8		31	30	22		20																									
6																															

NO. OF HYDROGEN ATOMS

NO. OF CARBON ATOMS

Figure 3. Populations (≧ 20) of molecular formula groups in the series C_xH_yO

groups which contain large numbers of compounds exhibit a regular pattern. This pattern may be used to predict the approximate number of compounds to be expected in any particular molecular formula group, an essential feature of the registry technique recently described by Lynch *et al.*[1]

ACKNOWLEDGMENT

We gratefully acknowledge a grant from the Office for Scientific and Technical Information, London. Part of this work was done by J. H. R. B. as a special study in partial fulfillment of the Diploma in Librarianship of the University of Sheffield.

LITERATURE CITED

(1) Lynch, M. F., J. Orton, and W. G. Town, *J. Chem. Soc., C*, **1969**, 1732–6.
(2) Bernays, P. M., Statistical data on chemical compounds, AD-615-488, 1965.
(3) Leighner, L. H., and D. P. Leiter, Jr. "A Statistical Analysis of the Structure Registry at Chemical Abstracts Service," Division of Chemical Literature, 154th Meeting, ACS, Chicago, Ill., September 1967.

The GREMAS System, an Integral Part of the IDC System for Chemical Documentation

SIGRID RÖSSLER AND ARTHUR KOLB
Farbenfabriken Bayer AG, Leverkusen, IDC Internationale Dokumentationsgesellschaft
für Chemie m.b.H., Frankfurt/Main, Germany

Received January 27, 1970

The Genealogical Retrieval by Magnetic Tape Storage (GREMAS) system and the potential it offers for searches are described. The input and retrieval procedures of the system are explained as well as the integration of the GREMAS system into the IDC system[1]—i.e., machine generation of the GREMAS coding from topological input and of the superimposed bit code from the GREMAS coding.

The GREMAS system serves to index low molecular organic compounds and compound classes. Essentially it is a fragment code which transcribes fragments of chemical structures into letter terms. These are then registered by a computer on magnetic tapes. Supplementary to the usual characteristics of the common fragment codes, the GREMAS system has additional features that enhance both the versatility and the hit rate in searching considerably. It was so designed as to give prime importance to a chemist's viewpoint; aspects of programming and machine processing became therefore subordinate. Consequently, the hierarchy of the system mirrors as closely as possible those principles of chemical classification that chemists use in their publications and inquiries. The rejection of too formalized tenets of classification forms one essential prerequisite for loss-free retrieval with minimal false drops.

The code is most selective in areas of chemistry comprising large numbers of compounds and having a high growth rate. But even a small percentage of false drops may become a problem with a file as large as the one which is formed by deep indexing of such a significant segment of the ever growing chemical literature. A still more specific system would then be desirable. By completely recording the topology of molecular structures—i.e., all atom-to-atom bonds—all the structural information of this compound is retained. But searching topological files takes so much computer time that it is economically unfeasable unless a preceding highly efficient and cheap search reduces the number of compounds to a minimum. Therefore, the IDC system employs a combination of the GREMAS search and a topological search. The GREMAS search precedes the topological retrieval. To reduce cost, the GREMAS search comprises a first screen—the superimposed bit code[2], a sort of "Abbreviated GREMAS"—which is computer-generated during input from each GREMAS term and is stored in front of the GREMAS file units on the magnetic tape.

The use of a polyhierarchical documentation system as sophisticated as the IDC system still remains expensive, but it is needed for selective retrieval. Although it is possible to look up specific compounds and parent structures in conventional card files and indexes, multidimensional processing is required for substructure searching. This can be handled adequately only by a computer when a large amount of literature has been stored. The GREMAS system allows selective substructure retrieval from a file containing currently about 900,000 compounds, including ones with alternative groups (Markush formulas). Moreover, reactions and types of reaction are searchable by themselves as well as in combination with inorganic reactants, catalysts, and nonstructural concepts.

THE GREMAS SYSTEM

The GREMAS system is a computerized storage and retrieval system for low molecular organic structures. It was developed by Farbwerke Hoechst AG[3] starting in