5. Use of MC (multicolumn) utility where appropriate (FORDEX, PERDEX, AUTDEX, JRNDEX) to produce final listing.

The storage of certain keys within the master Bibliographic File, the avoidance of step 4 for all but one index, and the use of locally written, fast utilities at steps 3 and 5 all serve to reduce computation time to a minimum. The use of step 5, where appropriate, also serves to reduce printed output to a minimum. The only high-level language involvement is in steps 1 and 4.

DISCUSSION

The six indexes described from an integrated cross-linked system which has a number of applications:

- As a stand-alone search tool providing data-base entry via the four major bibliographic information fields.
- As an adjunct to computer search techniques.[4] A rapid index scan yields accurate estimates of the number of "hits" to expect for a given query. Such knowledge may suggest query refinement to expand or diminish the scope of the search.
- As an aid to file maintenance and for spot checks on file consistency. The compound name and author indexes present material in an ordered form, particularly useful for visual scanning to detect spelling errors or lack of standardization. Such listings are always generated during the checkout of new input material.
- As part of our information dissemination program. The index system, on magnetic tape, is an integral part of regular data-base releases to 17 Affiliated Data Centres worldwide. All programs, except JRNDEX, are interfaced with typesetting software and contribute to MSD volumes[5,8] and, more recently, to the Organic Supplement of the NBS publication *Crystal Data*.[13] Finally a NAMDEX listing always accompanies the Current Awareness listing of new entries.

ACKNOWLEDGMENT

I wish to acknowledge the valuable programming contributions of Dr. J. R. Rodgers and Mrs. A. Doubleday, and the Staff of CCDC, particularly Drs. O. Kennard and D. G. Watson for advice and discussion. The Science Research Council and CCDC Affiliated Data Centres are thanked for financial support. All programs are written in Fortran IV for a 512K IBM 370/165 computer. Sort operations use the IBM Sort/Merge Package. The Staff of the University of Cambridge Computer Centre are thanked for advice and cooperation in the use of local utilities and implementations.

REFERENCES AND NOTES

(1) O. Kennard, D. G. Watson, F. H. Allen, W. D. S. Motherwell, W. G. Town, and J. R. Rodgers, *Chem. Br.*, **11**, 213 (1975).
(2) O. Kennard, D. G. Watson, and W. G. Town, *J. Chem. Doc.*, **12**, 14 (1972).
(3) F. H. Allen, O. Kennard, W. D. S. Motherwall, W. G. Town, and D. G. Watson, *J. Chem. Doc.*, **13**, 119 (1973).
(4) O. Kennard, F. H. Allen, M. D. Brice, T. W. A. Hummelink, W. D. S. Motherwell, J. R. Rodgers, and D. G. Watson, *Pure Appl. Chem.*, **49**, 1807 (1977).
(5) O. Kennard and D. G. Watson, "Molecular Structures and Dimensions", Vols. 1-3; with W. G. Town: Vols. 4, 5, Oosthoek, Utrecht, 1970, 1972, 1973, 1974; with F. H. Allen and S. M. Weeds: Vols. 6-10, Bohn, Scheltema, and Holkema, Utrecht, 1975, 1976, 1977, 1978, 1979.
(6) O. Kennard, D. G. Watson, F. H. Allen, N. W. Isaacs, W. D. S. Motherwell, R. C. Pettersen, and W. G. Town, "Molecular Structures and Dimensions", Vol. Al, Oosthoek, Utrecht, 1973.
(7) F. H. Allen, N. W. Isaacs, O. Kennard, W. D. S. Motherwell, R. C. Pettersen, W. G. Town, and D. G. Watson, *J. Chem. Doc.*, **13**, 211 (1973).
(8) O. Kennard, F. H. Allen and D. G. Watson, "Molecular Structures and Dimensions: Guide to the Literature 1935-1976", Bohn, Scheltema, and Holkema, Utrecht, 1977.
(9) H. Skolnik, *J. Chem. Inf. Comput. Sci.*, **16**, 187 (1976).
(10) F. H. Allen and W. G. Town, *J. Chem. Inf. Comput. Sci.*, **17**, 9 (1977).
(11) E. Garfield, *J. Chem. Doc.*, 3, 97 (1963).
(12) "The 1967 Volume Index", Chemical Abstracts Service, Columbus, Ohio, 1967.
(13) O. Kennard, D. G. Watson, J. R. Rodgers, and S. M. Weeds, "Crystal Data", 3rd ed, Vol. 3 (Organic Compounds), 1967-74, National Bureau of Standards, Washington D.C., 1979.

The Chemical Abstracts Service Chemical Registry System. VI. Substance-Related Statistics

ROBERT E. STOBAUGH
Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received October 24, 1979

Statistics on types of substances, ring systems, and elemental composition have been determined for the Chemical Abstracts Service Registry Structure File at different points in time. This paper reports these statistics and offers some comparisons to show the various shifts in file characteristics.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system that uniquely identifies chemical substances on the basis of their molecular structure. The design, content, and functions of the Registry System have been described in detail in previous papers.[1-5] In addition, the function of the system as an interfile linking agent for information resources has also been described.[6]

The computer-readable structure records that make up the Registry files are basically records of the atoms and bonds present in the molecular structure of the substances. They represent the ring systems that are present, the substituents attached to the rings, and any substituents that link two or more rings. From these structure records, statistics can be obtained routinely and with little difficulty for analyses of elemental composition and ring characteristics. These statistics, along with those for types of structures, are presented in this paper.

Since December 1978, statistics have been determined for the various classes of substances in the CAS Chemical Registry System files. The tables in this paper present a comparison of cumulative occurrence data concerning ring graphs for the years 1974 and 1978 and ring systems for 1974, 1976, and 1978. Also compared are the cumulative occurrence data for elemental composition for the years 1967, 1974, and 1979. Tables report the percentage increase from 1974 to 1979 for the occurrence of elements and for substances containing the given elements. Similarly, statistics are provided for the

0095-2338/80/1620-0076$01.00/0

Table I. CAS Chemical Registry Coverage (December 1978)

Types of Substance	Number
Fully-defined substances	3,622,448
Incompletely-defined substances	63,807
Polymers	160,845
Coordination compounds	288,778
Alloys	82,715
Mixtures	10,822
Minerals	1,416
Radical ions	7,584
Ring parents	45,764
TOTAL	4,284,159

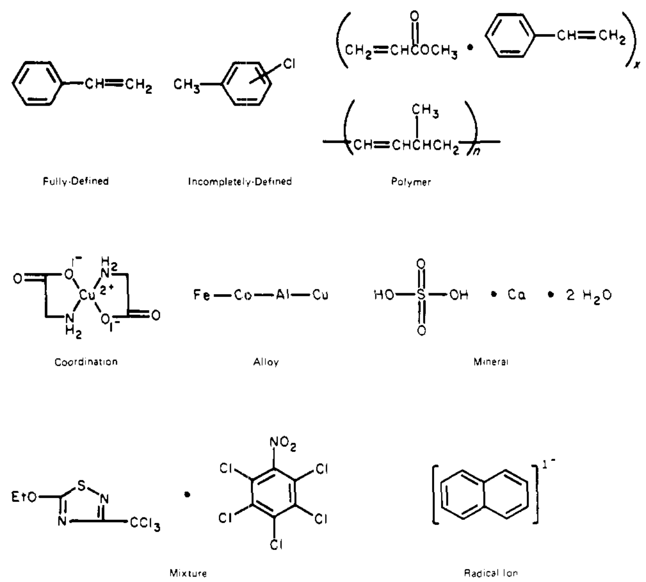


Figure 1. Types of substances.

percentage increase for ring graphs and ring systems.

TYPES OF SUBSTANCES

A list of the numbers of registered substances for various classes of chemical substances as of December 1978 is provided in Table I. These numbers are for those substances registered by machine processing and do not include those registered by manual techniques.[7] This is by no means an exhaustive classification of chemical substances. The only classes listed are those for which statistics are routinely obtained during operation of the CAS Chemical Registry System. The classes are mutually exclusive; e.g., a coordination compound is not also counted as fully defined, or incompletely defined, or a polymer, even though it is also a member of at least one of these classes.

Examples of these classes are shown in Figure 1. More detailed descriptions of the classes have previously been published,[1] except for the class "ring parent". A ring parent is the particular bond variation of a ring system chosen as the representative of the family of index parents which is illustrated in the "Parent Compound Handbook"[8] and the CA Chemical Substance Index to show ring-system numbering. Even when the ring parent itself does not exist, it is entered into the CAS Chemical Registry System. Figure 2 provides examples of ring parents.

Figure 2. Ring parents.

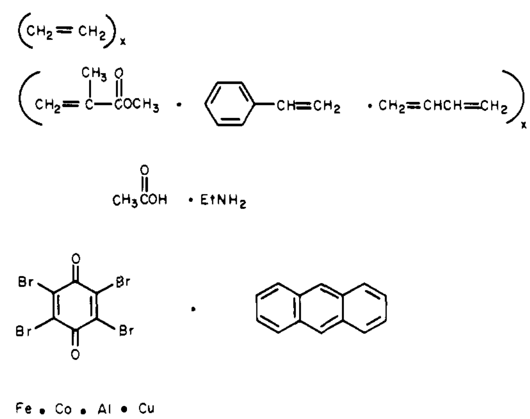


Figure 3. Expressions.

Table II. Expression Statistics (December 1978)

Expressions	528,913
With one component	39,319
With two components	374,233
With three components	51,379
With four or more components	63,622
Components/expression - - average	2.4
Components/expression - - highest	19

STRUCTURAL CHARACTERISTICS

Chemical substances are recorded in the CAS Chemical Registry System in terms of components,[1] a component being a set of contiguous atoms. For some substances more than one component is necessary for adequate representation, for example, salts of acids, salts of bases, polymers, mixtures, and other types. In the CAS Chemical Registry System such a combination of components is called an "expression". Figure 3 gives some examples. About 12.2% of the substances recorded are made up of expressions, as shown in Table II. From 1 to 19 components can be present in expressions; the average number is 2.4. An expression may consist of only one component, if additional information such as the polymer subscript x is present. Expressions containing two components are by far the most common. Those containing more than five are almost always alloys.

The large majority, 87.9%, of chemical substances are represented by a single component, and these may either be acyclic or cyclic. Table III lists statistics for these single component substances. A relatively small percentage (12.5%) contain no rings at all. The number of atoms in these acyclic substances averages approximately 15. Of the cyclic substances, almost half (48.0%) contain one ring system and another third (31.8%) contain two ring systems. A ring system is defined as being any cyclic arrangement of atoms and bonds, for example, cyclopentane, naphthalene, pyrrole, or perylene.

A basic design feature of the CAS Chemical Registry System is that the ring systems present in a structure are recognized during the registration process.[1] The systems are stored in the structural record of a given substance only as an identifying number which links the structure record to a file of ring systems. In this file, ring systems are recorded as composites of the ring graphs, or basic patterns; as graph-node

Table III. Component Statistics (December 1978)

Components	3,755,246
Acyclic	465,030
Atoms/acyclic component - - average	14.9
Cyclic	3,290,216
With one ring system	1,571,517
With two ring systems	1,046,907
With three ring systems	411,852
With four or more systems	259,940
Ring systems/cyclic component - - average	1.8

Table IV. Ring Statistics

	Total	Average	Highest
Ring graphs	22,463		
Rings/ring graph		7.2	4,751
Ring-node sets	85,351		
Sets/ring graph		3.7	1,421
Coordination ring node sets	36,146		
Ring-node-bond sets	162,859		
Sets/ring node variant		1.9	322

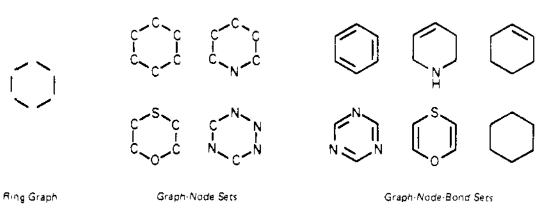


Figure 4. Examples of ring graph-node-bond variations.

(atom) variations, in which the atoms have specific identities; and as graph-node-bond variations for fully specified systems. Examples of these are given in Figure 4. Statistics for the ring system constituents, given in Table IV, show that in the total Registry file of over 4 million substances, there are only 22 463 basic ring graphs and 162 859 different ring systems. The number of rings per graph ranges from 1 to 4751, the average number being 7.2. For each graph, there is an average of 3.7 different sets of specified atoms, the range being from 1 to 1421. For each graph-node variant, there are almost 2 (1.9) specific bond variations, ranging from 1 to 321.

Ring systems are very common in chemical substances. Not only do 3 290 216 substances contain them, but there are 5 922 389 total occurrences, since many substances contain more than one ring system. Of the total occurrences, 3 269 266 (55%) are phenyl rings, both unsubstituted and substituted. There are 200 ring graphs that account for over 96% of the ring systems' occurrences. These 200 ring graphs are categorized as listed in Table V. More fused ring systems are present than any other type of ring system, but there is a wide variety of structural types including the complex of metallocenes and the polyhedral cage typical of some carboranes.

The 20 most frequently occurring ring graphs, with the number of occurrences given as of March 1974 and December

Table V. 200 Most Frequent Ring Graphs

Type	Number
Fused	98
Spiro-fused	26
Bridged-fused	25
Bridged	17
Single	15
Spiro	13
Macro	3
π-complex	2
Polyhedral	1

1978, are pictured in Table VI. It is not surprising that the six-membered graph, typical of benzenes, pyridines, pyrans, morpholines, pyrimidines, etc., is the most frequently occurring. The remaining three of the most frequent four ring graphs include the five-membered graph, the basis for cyclopentanes, pyrroles, thiazoles, furans, etc.; the two ortho-fused six-membered graphs, typical of naphthalenes, quinolines, benzopyrans, etc.; and the ortho-fused six- and five-membered graphs, basic to indenes, indoles, benzothiophenes, etc. Altogether these top-ranking four types of graphs comprise the majority of ring graphs.

In considering specific ring systems, the benzene ring is by far the most common, with pyridine second, cyclohexane third, naphthalene fourth, and piperidine fifth. Table VII shows the 20 most frequently occurring ring systems with frequencies given as of March 1974, June 1976, and December 1978. The top three have remained in the same order consistently, while there have been some variations in the order of the less frequently occurring systems. Six-membered rings, single and fused, account for over 69% of all rings.

The total number of ring graphs showed a 55.4% increase from 1974 to 1978. Examination of the percentage increase in the number of occurrences of individual ring graphs from 1974 to 1978 shows that, of the 20 most common, the bicyclo[4.2.0] graph increased the most (222%), followed by the bicyclo[3.2.0] graph (131%), and the bicyclo[3.3.0] graph (116%), with the tetracyclic graph typical of steroids showing the least increase (38.6%) in frequency. Such figures indicate that the 1974 to 1978 period was one of high publishing activity for research in the fields of cephalosporins, penicillins, and prostaglandins, but of relatively low activity for steroids.

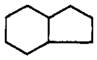
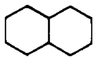
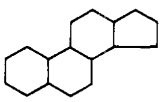
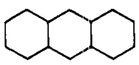
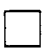
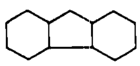
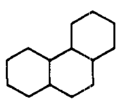
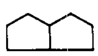
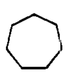
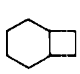
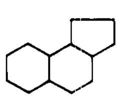
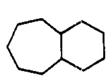
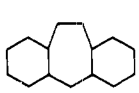
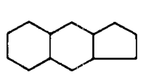
Figures on the percentage increase in occurrence of ring systems support this indication for steroids, since the only steroid-type ring system appearing in the 20 most frequently occurring ring graphs shows an increase of only 42%, the lowest of the top 20. There is no one specific bicyclo[4.2.0], bicyclo[3.2.0], or bicyclo[3.3.0] system appearing in the first 20 ring systems. Hence, the increase in occurrence of those three ring graphs probably involves a number of ring systems, but the cephalosporins, penicillins, and prostaglandin-related compounds must certainly play a prominent role. The cyclopentane ring system shows the largest increase (145.8%) in number of occurrences from 1974 to 1978, followed by the tetrahydropyran system (121.4%), and the tetrahydrofuran system (101.4%).

ELEMENT STATISTICS

Statistics have been obtained at varying time intervals on the occurrence of the elements in the CAS Registry Structure File. Table VIII shows the number of substances containing

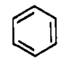
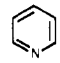
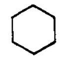
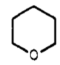
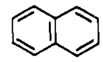
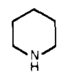
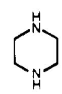
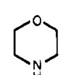
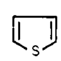
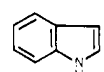
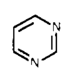
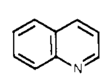
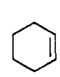
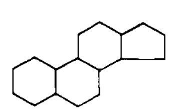
J. Chem. Inf. Comput. Sci., *Vol. 20, No. 2, 1980* **79**

Table VI. Ring Graph Frequency

Ring Graph	Occurrences 1974	1978
	2,362,044	4,143,118
	368,894	689,176
	202,188	358,567
	210,128	342,957
	67,395	93,378
	47,116	85,826
	45,175	70,063
	21,325	39,303
	22,954	39,161
	21,282	35,109
	15,150	32,784
	16,299	26,560
	16,239	26,321
	13,500	23,447
	7,194	23,189
	13,572	23,124
	11,372	19,298
	8,339	19,218
	11,018	16,421
	8,729	15,739

a given element and the percent of the total as of 1967,[9] 1974, and 1979, and Table IX lists the analogous data for occurrences (number of atoms) of the elements. The 1967 statistics are included for historical more than comparative purposes.

Table VII. Frequently Occurring Ring Systems

Ring System	Occurrences 1974	1976	1978
	1,861,106	2,577,641	3,269,266
	95,086	130,190	161,097
	80,004	112,132	142,050
	44,935	69,716	99,482
	61,509	78,793	95,295
	57,048	74,685	93,236
	37,767	56,355	76,062
	40,285	56,110	75,046
	27,794	39,152	48,643
	27,959	37,565	46,152
	27,116	36,751	44,873
	22,549	32,128	41,141
	23,139	32,739	41,068
	25,531	33,743	40,829
	16,366	26,527	40,224
	25,087	32,564	39,564
	17,130	24,570	33,304
	17,919	25,801	33,147
	22,323	27,236	31,693
	16,912	24,496	31,554

At that time the CAS Chemical Registry System had been operating for not quite $2^1/_2$ years and contained the substances

Table VIII. Elemental Composition Statistics by Substance

	5/31/67		2/28/74		2/18/79	
Element Symbol	No. of Substances	Percent	No. of Substances	Percent	No. of Substances	Percent
Ac			62	0.002275	106	0.002328
Ag	88	0.014756	6,336	0.232474	12,162	0.267151
Al	909	0.152422	19,436	0.713127	37,902	0.832558
Am	2	0.000335	332	0.012181	451	0.009906
Ar	2	0.000335	127	0.004660	200	0.004393
As	2,165	0.363030	15,162	0.556309	25,162	0.552710
At	2	0.000335	86	0.003155	160	0.003514
Au	144	0.024146	3,117	0.114366	6,541	0.145876
B	5,417	0.908333	36,057	1.322970	58,428	1.283433
Ba	39	0.006539	4,282	0.157111	7,035	0.154531
Be	76	0.012744	2,669	0.097928	4,174	0.091686
Bi	186	0.031189	3,002	0.110146	5,367	0.117891
Bk			53	0.001945	107	0.002350
Br	24,378	4.087750	147,014	5.394095	243,199	5.342127
C	594,706	99.676353	2,630,958	96.532551	4,293,917	94.320495
Ca	72	0.012073	6,839	0.250930	11,770	0.258540
Cd	120	0.020113	6,266	0.229906	10,445	0.229435
Ce	8	0.001341	2,412	0.088499	4,363	0.095837
Cf			76	0.002789	134	0.002943
Cl	83,432	13.990043	544,891	19.992610	894,992	19.659459
Cm			130	0.004770	214	0.004700
Co	179	0.030015	38,157	1.400014	67,243	1.477064
Cr	136	0.022805	22,740	0.834354	49,691	1.091516
Cs	25	0.004192	3,789	0.139022	6,275	0.137837
Cu	251	0.042088	32,635	1.197412	61,081	1.341709
D	2,806	0.470516	15,977	0.586213	30,028	0.659597
Dy	7	0.001173	1,334	0.048946	2,378	0.052235
Er	10	0.001676	1,575	0.057788	2,655	0.058319
Es			33	0.001211	88	0.001933
Eu	8	0.001341	2,043	0.074960	3,308	0.072663
F	58,687	9.840752	166,386	6.104873	265,714	5.836693
Fe	146	0.024481	40,322	1.479456	90,688	1.992059
Fm			42	0.001541	77	0.001691
Fr			45	0.001651	78	0.001713
Ga	104	0.017439	3,252	0.119319	5,594	0.122878
Gd	9	0.001509	1,888	0.069273	3,447	0.075717
Ge	1,306	0.218992	8,677	0.318368	14,030	0.308184
H	596,347	99.996646	2,636,306	96.728775	4,289,454	94.222461
He			107	0.003926	156	0.003425
Hf	8	0.001341	1,252	0.045937	2,455	0.053926
Hg	1,769	0.296629	11,601	0.425653	18,273	0.401386
Ho	3	0.000503	1,214	0.044543	2,017	0.044305
I	6,957	1.166563	64,849	2.379376	99,516	2.185975
In	45	0.007546	3,279	0.120310	5,817	0.127776
Ir	16	0.002683	4,717	0.173072	8,472	0.186096
K	161	0.026997	17,433	0.639635	28,174	0.618872
Kr			184	0.006751	290	0.006370
La	11	0.001844	2,843	0.104313	4,847	0.106469
Li	298	0.049969	7,786	0.285676	14,496	0.318420
Lr			14	0.000514	53	0.001164
Lu	3	0.000503	931	0.034159	1,521	0.033410
Md			25	0.000917	58	0.001274
Mg	320	0.053658	8,726	0.320166	16,277	0.357541
Mn	66	0.011067	19,931	0.731289	48,062	1.055733
Mo	29	0.004863	14,476	0.531139	34,336	0.754227
N	383,050	64.230583	1,751,974	64.281725	2,872,142	63.089681
Na	187	0.031356	41,217	1.512295	71,326	1.566752
Nb	50	0.008384	5,328	0.195490	11,441	0.251313
Nd	10	0.001677	2,435	0.089343	4,016	0.088215
Ne			100	0.003669	135	0.002965
Ni	117	0.019619	34,294	1.258282	69,436	1.525236
No			23	0.000844	64	0.001405
Np			561	0.020584	886	0.019461
O	492,746	82.624625	2,215,027	81.271616	3,630,656	79.751256
Os	47	0.007881	2,000	0.073382	4,062	0.089226
P	28,044	4.702473	169,677	6.225623	292,133	6.417014
Pa			356	0.013062	441	0.009687
Pb	404	0.067743	6,707	0.246087	11,351	0.249336
Pd	36	0.006036	8,564	0.314222	17,305	0.380122
Pm			143	0.005247	205	0.004503
Po	62	0.010396	162	0.005944	241	0.005293
Pr	9	0.001509	1,894	0.069493	3,181	0.069874
Pt	172	0.028841	12,378	0.454162	24,745	0.543550
Pu	3	0.000503	656	0.024069	980	0.021526
Ra	4	0.000671	72	0.002642	119	0.002613
Rb	14	0.002347	2,485	0.091214	4,140	0.090939
Re	13	0.002180	3,606	0.132308	6,526	0.143350
Rh	3	0.000503	6,906	0.253388	14,108	0.309897
Rn			54	0.001981	81	0.001779
Ru	7	0.001174	4,746	0.174136	9,724	0.213598
S	118,536	19.876351	569,805	20.905731	967,434	21.250726
Sb	881	0.147728	8,442	0.309746	14,741	0.323801
Sc	16	0.002683	1,513	0.055514	2,163	0.047512
Se	2,116	0.354815	14,765	0.541743	25,251	0.554655
Si	8,628	1.446760	56,152	2.060285	125,165	2.749383
Sm	46	0.007713	2,068	0.075877	3,483	0.076507
Sn	2,857	0.479067	21,605	0.792710	36,652	0.805100
Sr	25	0.004192	2,325	0.085307	3,863	0.084854
T	579	0.097088	2,313	0.084866	4,059	0.089379
Ta	38	0.006372	2,971	0.109009	6,024	0.132323
Tb	5	0.000838	1,190	0.043662	2,002	0.043976
Tc	4	0.000671	345	0.012658	650	0.014277
Te	225	0.037728	3,506	0.128639	6,153	0.135157
Th	48	0.008049	1,966	0.072135	3,053	0.067062
Ti	407	0.068247	12,402	0.455042	24,710	0.542781
Tl	129	0.021631	3,467	0.127208	5,846	0.128413
Tm	3	0.000503	829	0.030417	1,361	0.029895
U	137	0.022972	5,799	0.212771	9,675	0.212521
V	145	0.024314	8,705	0.319395	17,437	0.383022
W	57	0.009558	9,255	0.297564	18,122	0.398069
Xe			422	0.015484	574	0.012608
Y	9	0.001509	2,192	0.080427	3,985	0.087534
Yb	5	0.000838	1,320	0.048432	2,233	0.049050
Zn	595	0.099771	14,703	0.539468	25,030	0.549810
Zr	57	0.009558	4,914	0.180300	9,703	0.213136

Table IX. Elemental Composition Statistics by Occurrence

	5/31/67		2/28/74		2/18/79	
Element Symbol	No. of Occurrences	Percent	No. of Occurrences	Percent	No. of Occurrences	Percent
Ac			64	.000054	108	.000053
Ag	91	.0003	9,129	.007725	15,448	.007703
Al	947	.0039	29,753	.025177	50,972	.025417
Am	2		368	.000311	504	.000251
Ar	2		173	.000146	243	.000121
As	2,496	.0102	25,610	.021672	42,531	.021208
At	2		93	.000079	167	.000083
Au	144	.0006	4,530	.003833	8,092	.004035
B	7,279	.0300	115,237	.097516	182,493	.091002
Ba	39	.0002	5,717	.004838	7,618	.003798
Be	77	.0003	4,664	.003947	5,743	.002863
Bi	193	.0008	5,101	.004317	9,314	.004644
Bk			56	.000047	115	.000057
Br	34,255	.1413	228,223	.193126	377,743	.188365
C	9,400,518	38.7785	44,959,807	38.045697	76,131,591	37.963829
Ca	73	.0003	9,915	.008390	14,051	.007006
Cd	120	.0004	7,653	.006476	12,190	.006078
Ce	8		3,384	.002864	5,458	.002721
Cf			79	.000067	147	.000073
Cl	149,248	.6157	974,531	.824663	1,587,485	.791616
Cm			163	.000138	255	.000127
Co	198	.0008	46,754	.039564	79,123	.039455
Cr	159	.0006	26,216	.022184	55,462	.027656
Cs	26	.0001	4,423	.003743	6,602	.003292
Cu	262	.0010	39,908	.033771	72,696	.036250
D	7,833	.0323	52,560	.044477	99,610	.049671
Dy	7		2,043	.001729	3,263	.001627
Er	10		2,764	.002339	3,455	.001722
Es			35	.000030	90	.000044
Eu	8		2,473	.002093	4,195	.002091
F	211,854	.8730	723,577	.612302	1,181,628	.589231
Fe	149	.0006	51,396	.043492	108,053	.053881
Fm			44	.000037	79	.000039
Fr			49	.000041	80	.000039
Ga	108	.0004	5,290	.004476	7,872	.003925
Gd	9		3,014	.002550	4,820	.002403
Ge	1,563	.0064	13,219	.011186	20,331	.010138
H	11,547,701	47.6360	55,600,781	47.050257	94,297,072	47.022240
He			137	.000116	203	.000101
Hf	8		1,835	.001553	3,032	.001511
Hg	1,974	.0081	14,071	.011907	22,044	.010992
Ho	3		1,844	.001560	2,770	.001381
I	10,429	.0430	94,153	.079674	146,121	.072864
In	47		4,406	.003728	7,187	.003583
Ir	16		5,322	.004504	9,682	.004828
K	180	.0007	19,439	.016450	29,139	.014530
Kr			191	.000162	308	.000153
La	11		4,590	.003884	7,212	.003596
Li	347	.0014	9,971	.008438	16,536	.008245
Lr			16	.000014	55	.000027
Lu	3		1,248	.001056	1,921	.000957
Md			27	.000023	60	.000029
Mg	328	.0013	12,155	.010286	18,916	.009432
Mn	69	.0002	22,979	.019445	52,376	.026117
Mo	29	.0001	24,867	.021043	53,268	.026562
N	917,258	3.7838	4,674,936	3.956004	8,040,072	4.009267
Na	215	.0009	44,699	.037825	73,581	.036691
Nb	50	.0002	11,393	.009641	20,772	.010358
Nd	10		3,652	.003090	5,556	.002770
Ne			121	.000102	161	.000080
Ni	119	.0004	39,706	.033600	77,127	.038460
No			27	.000023	66	.000032
Np			623	.000527	988	.000492
O	1,715,545	7.0769	8,702,367	7.364080	14,932,265	7.446133
Os	47	.0002	2,669	.002259	5,864	.002924
P	35,706	.1473	260,947	.220817	466,904	.232826
Pa			416	.000352	509	.000253
Pb	470	.0019	9,529	.008064	15,205	.007582
Pd	37	.0001	10,740	.009088	20,936	.010439
Pm			156	.000132	229	.000114
Po	62	.0002	165	.000140	245	.000122
Pr	9		2,873	.002431	4,276	.002132
Pt	174	.0007	14,426	.012208	28,255	.014089
Pu	3		891	.000754	1,088	.000542
Ra	4		72	.000061	119	.000059
Rb	14		3,061	.002590	4,445	.002216
Re	13		4,980	.004214	8,874	.004425
Rh	3		8,810	.007455	17,881	.008916
Rn			54	.000046	81	.000040
Ru	7		5,987	.005066	12,227	.006097
S	166,664	.6875	881,655	.746070	1,500,099	.748040
Sb	967	.0040	10,787	.009128	19,600	.009773
Sc	16		1,986	.001681	2,965	.001478
Se	2,534	.0104	25,105	.021244	41,282	.020585
Si	16,362	.0675	120,257	.101763	208,384	.103912
Sm	46	.0002	2,903	.002457	4,699	.002343
Sn	3,496	.0144	26,774	.022657	43,476	.021679
Sr	25	.0001	3,684	.003117	4,518	.002252
T	857	.0035	3,486	.002950	6,458	.003220
Ta	38	.0001	5,653	.004784	10,251	.005111
Tb	5		1,740	.001472	2,659	.001340
Tc	4		424	.000360	749	.000373
Te	231	.0009	6,889	.005830	10,881	.005425
Th	48	.0002	2,349	.001988	3,587	.001788
Ti	478	.0019	17,361	.014691	29,962	.014940
Tl	140	.0006	4,193	.003548	6,888	.003434
Tm	3		1,170	.000990	1,760	.000877
U	141	.0006	7,512	.006357	12,681	.006323
V	148	.0006	14,492	.012263	26,794	.013361
W	59	.0002	134,149	.020113	44,549	.022214
Xe			462	.000391	643	.000320
Y	9		3,442	.002913	5,700	.002842
Yb	5		1,854	.001569	2,932	.001462
Zn	613	.0025	19,132	.016190	28,311	.014117
Zr	57	.0002	6,747	.005709	12,063	.006015

indexed for almost five volumes of *Chemical Abstracts*, the total collection of fluorine-containing compounds from 1907 to date, and some CAS internal working files. Organic compounds only were being processed, and these did not include polymers, coordination compounds, or incompletely defined substances; hence the file as of 31 May 1967 was not representative of the entire chemical field. Statistics of 1974 and 1979, however, have been obtained from Registry files that encompass all types of chemical substances and are fully representative.

Examination of the statistics of Tables VIII and IX corroborates some well-known facts and reveals others not so evident. Hydrogen (not including its isotopes, deuterium and tritium) occurs in more substances than any other element, followed closely by carbon. In succession, in lesser percentage for all these samples, are oxygen, nitrogen, sulfur, and chlorine. Table X shows the first 20 most commonly occurring elements in substances. After the first six, which are identical, there are variations, but relatively minor ones between the 1974 and 1979 samples. The 1967 rankings show the bias toward

Table X. Twenty Most Frequently Occurring Elements According to Number of Substances

1967		1974		1979	
H	99.99%	H	96.73%	C	94.32%
C	99.68	C	96.53	H	94.22
O	82.62	O	81.27	O	79.75
N	64.23	N	64.28	N	63.09
S	19.88	S	20.91	S	21.25
Cl	13.99	Cl	19.99	Cl	19.66
F	9.84	P	6.23	P	6.42
P	4.70	F	6.10	F	5.84
Br	4.09	Br	5.39	Br	5.34
Si	1.45	Si	2.43	Si	2.75
I	1.17	I	2.34	I	2.19
B	0.91	Na	1.51	Fe	1.99
Sn	0.48	Fe	1.48	Na	1.57
D	0.47	Co	1.40	Ni	1.53
As	0.36	B	1.32	Co	1.48
Se	0.35	Ni	1.26	Cu	1.34
Hg	0.30	Cu	1.20	B	1.28
Ge	0.22	Cr	0.83	Cr	1.09
Al	0.152	Sn	0.79	Mn	1.06
Sb	0.148	Mn	0.73	Al	0.83

Table XI. Twenty Most Frequently Occurring Elements According to Number of Atoms

1967		1974		1979	
H	47.64%	H	47.05%	H	47.02%
C	38.79	C	38.05	C	37.96
O	7.08	O	7.36	O	7.45
N	3.78	N	3.96	N	4.01
F	0.87	Cl	0.82	Cl	0.79
S	0.69	S	0.75	S	0.75
Cl	0.62	F	0.61	F	0.59
P	0.15	P	0.22	P	0.23
Br	0.14	Br	0.19	Br	0.19
Si	0.07	Si	0.10	Si	0.10
I	0.04	B	0.098	B	0.093
D	0.032	I	0.080	I	0.074
B	0.030	D	0.044	Fe	0.052
Sn	0.014	Fe	0.044	D	0.050
Se	0.010	Co	0.040	Co	0.039
As	0.010	Cu	0.034	Ni	0.038
Hg	0.008	Ni	0.034	Cu	0.036
Ge	0.006	Al	0.025	Cr	0.028
Al	0.004	Sn	0.023	Al	0.025
Sb	0.004	Cr	0.022	Sn	0.022

Table XII. Registry Structure File Statistics

	1967	1974	1979
No. of substances	596,367	2,725,462	4,552,475
No. of atoms	24,241,534	118,173,172	200,537,175

fluorine, for the reason stated previously.

Table XI shows the 20 highest occurring elements in terms of total numbers of atoms, rather than substances. Hydrogen, carbon, oxygen, and nitrogen, in that order, head the list as illustrated in Table X. Below these, there is some variation, again, only minor, between the 1974 and 1979 samples.

Table XII lists the number of substances and the number of total atoms in the three samples of the structure file. The percentage increase in number of substances from the 1974 to 1979 sample is 67.0%, for the number of atoms, 69.7%. For individual elements, the percentage increase (Table XIII) of substances and of occurrences (atoms) varied widely from 22.4% (protactinium) to 278.6% (lawrencium) for substances, and from 22.1% (protactinium) to 243% (lawrencium) for occurrence. Several of the actinide elements, berkelium, einsteinium, lawrencium, nobelium, and mendelevium, showed very high percentage increases in substances containing them and in total occurrence in the 1974–1979 period. However,

Table XIII. Elemental Composition Increases 1974 to 1979

	Substances No.	%	Occurrences No.	%
Ac	44	71.0	44	68.8
Ag	5,826	92.0	6,319	69.2
Al	18,465	95.0	21,219	71.3
Am	119	35.8	136	37.0
Ar	73	57.5	70	40.5
As	10,000	66.0	16,921	66.1
At	74	86.0	74	80.0
Au	3,524	113.1	3,562	78.6
B	22,371	62.0	67,256	58.4
Ba	2,753	64.3	1,901	33.3
Be	1,505	55.4	1,079	23.1
Bi	2,365	78.8	4,213	82.6
Bk	54	101.9	59	105.4
Br	95,185	65.4	149,520	65.5
C	1,662,959	63.2	31,171,184	69.3
Ca	4,931	72.1	4,136	41.7
Cd	4,179	66.7	4,537	59.3
Ce	1,951	80.8	2,074	61.3
Cf	58	76.3	68	86.1
Cl	350,101	64.3	612,955	62.9
Cm	84	64.6	92	56.4
Co	29,086	76.2	32,369	69.2
Cr	26,951	118.5	29,246	111.6
Cs	2,486	65.6	2,179	49.3
Cu	28,446	87.2	32,788	82.2
D	14,031	87.1	47,041	89.5
Dy	1044	78.3	1220	59.7
Er	1080	68.6	691	25.0
Es	55	166.7	55	157.1
Eu	1265	61.2	1,722	69.6
F	99,328	59.7	458,051	63.3
Fe	50,366	124.9	56,657	110.2
Fm	35	83.3	35	79.5
Fr	33	73.3	31	63.3
Ga	2,342	72.0	2,582	48.8
Gd	1,559	82.6	1,806	59.9
Ge	5,353	61.7	7,112	53.8
H	1,653,148	52.7	38,696,291	69.6
He	49	45.8	66	48.2
Hf	1,203	96.1	1,197	65.2
Hg	6,672	57.5	7,973	56.7
Ho	803	66.1	926	50.2
I	34,667	53.5	51,968	55.2
In	2,538	77.4	2,781	63.1
Ir	3,755	79.6	4,360	81.9
K	10,741	61.6	9,700	49.9
Kr	106	57.6	117	61.3
La	2,004	70.5	2,622	57.1
Li	6,710	86.2	6,565	65.8
Lr	39	278.6	39	243.8
Lu	590	63.4	673	53.9
Md	33	132.0	33	122.2
Mg	7,551	86.5	6,761	55.6
Mn	28,131	141.1	29,397	127.9
Mo	19,860	137.2	28,401	114.2
N	1,120,168	63.9	3,365,136	72.0
Na	30,109	73.1	28,881	64.6
Nb	6,113	114.7	9,379	82.3
Nd	1,581	64.9	1,904	52.1
Ne	35	35.0	40	33.3
Ni	35,142	102.5	37,421	94.3
No	41	175.3	39	144.4
Np	325	57.9	365	58.6
O	1,415,629	63.9	6,229,898	71.6
Os	2,062	103.1	3,195	119.7
P	122,458	72.2	205,597	78.9
Pa	85	23.9	93	22.1
Pb	4,644	69.2	5,675	59.6
Pd	8,741	102.1	10,195	94.9
Pm	62	43.4	73	46.8
Po	79	48.8	80	48.5
Pr	1,287	68.0	1,403	48.8
Pt	12,367	99.9	13,829	95.9
Pu	324	49.4	197	22.1
Ra	47	65.3	47	65.3
Rb	1,654	66.5	1,384	45.2
Re	2,920	81.0	3,894	75.2
Rh	7,202	104.3	9,071	103.0
Rn	27	50.0	27	50.0
Ru	4,978	104.9	6,240	104.2
S	397,629	69.8	618,444	70.1
Sb	6,299	74.6	8,813	81.7
Sc	650	43.0	979	49.3
Se	10,486	71.0	15,177	64.4
Si	59,013	89.2	88,127	73.3
Sm	1,415	65.4	1,796	61.9
Sn	15,047	69.6	16,702	60.4
Sr	1,538	66.2	834	22.6
T	1,756	75.9	2,927	85.3
Ta	3,053	102.8	4,598	81.3
Tb	812	68.2	949	54.5
Tc	305	88.4	325	76.7
Te	2,647	75.5	3,992	57.9
Th	1,087	55.3	1,238	52.7
Ti	12,308	99.2	12,601	72.6
Tl	2,379	68.5	2,694	64.2
Tm	532	64.2	590	50.4
U	3,876	66.8	5,169	68.8
V	8,732	100.3	12,302	84.9
W	7,696	94.9	15,516	65.3
Xe	152	36.0	181	39.2
Y	1,793	81.8	2,258	65.6
Yb	913	69.2	1,078	58.1
Zn	10,327	70.2	9,179	48.0
Zr	4,789	97.9	5,316	78.8

substances containing several more common elements showed very high percentage increases: gold (113.1%), chromium (118.5), iron (124.9%), manganese (141.1%), molybdenum (137.2%), rhodium (104.3%), ruthenium (104.3%), osmium (103.1%), and nickel (102.5%). For total occurrence (atoms), a similar list is obtained: chromium (111.6%), iron (110.2%), manganese (127.9%), molybdenum (114.2%), osmium (119.7%), rhodium (103.0%), and ruthenium (104.2%).

The statistics reported here show variations in ring and element characteristics over relatively short time periods. The absolute values will increase with time, but the percentage of the total file will probably show little change. Percentage increase over time may show substantial variation, depending on the substances reported in the literature.

CONCLUSION

These statistics have been presented for whatever use may be made of them. There has been limited use in the past since the statistics on frequency have not been widely available. Frequency figures on ring system occurrence are used in the definition and internal processing of certain screens in the CAS Online Substructure Search System presently under development.[10] Statistics on elemental occurrence have been used in the development of molecular formula screens in the substructure search system based on *Chemical Abstracts* index nomenclature.[11,12]

Applications of pattern recognition techniques to the study of structure–activity relationships have employed structural features involving elemental composition and ring systems among others.[13,14] These applications involve the presence of the feature rather than any frequency data. However, frequency figures on ring systems or elemental composition might indicate a particular class of substance as a field for investigation.

It is hoped that the statistics in this paper will provoke new ideas and therefore stimulate research in chemistry, chemometrics, and information science.

ACKNOWLEDGMENT

The development of the CAS Chemical Registry System was substantially supported by the National Science Foundation. Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

Supplementary Material Available: (1) The cumulative occurrence statistics for the 199 most frequently occurring ring graphs in the CAS Chemical Registry System for the years 1974 and 1978 (25 pages). (2) The cumulative occurrence statistics for the 198 most frequently occurring ring systems in the CAS Chemical Registry System for the years 1974, 1976, and 1978 (20 pages). Ordering information is given on any current masthead page. Copies of these statistics may also be obtained in printed form at the following address: Marketing Communications Department, Chemical Abstracts Service, 2540 Olentangy River Road, P. O. Box 3012, Columbus, Ohio 43210.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. I. General Design". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–124.
(2) Freeland, R. G.; Funk, S. J.; O'Korn, L. J.; Wilson, G. A. "The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molecular Formula", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94–98.
(3) Blackwood, J. E.; Elliott, P. S.; Stobaugh, R. E.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3–8.
(4) Vander Stouw, G. G.; Gustafson, C. R.; Rule, J. D.; Watson, C. E. "The Chemical Abstracts Service Chemical Registry System. IV. Use of the Registry System to Support the Preparation of Index Nomenclature", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 213–18.
(5) Zamora, A.; Dayton, D. L. "The Chemical Abstracts Service Chemical Registry System. V. Structure Input and Editing", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 219–22.
(6) Myers, D. C.; Rathbun, J. A.; Tate, F. A.; Weisgerber, D. W. "Bridging and Linking the Information Resources", *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 16–19.
(7) Moosemiller, J. P.; Ryan, A. W.; Stobaugh R. E. "The Chemical Abstracts Service Chemical Registry System. VIII. Manual Registration", *J. Chem. Inf. Comput. Sci.*, following paper in this issue.
(8) Blake, J. E.; Brown, S. M.; Ebe, T.; Goodson, L. A.; Skevington, J. H.; Watson, C. E. "Parent Compound Handbook: The Successor to the Ring Index", *J. Chem. Inf. Comput. Sci.*, in preparation.
(9) Leiter, D. P.; Leighner, L. "A Statistical Analysis of the Structure Registry at Chemical Abstracts Service", presented at the 154th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 1967.
(10) Fisanick, W.; Haines, R. C. "An Online Substructure Search System. I. Screens and Screening Techniques", presented to the 176th National Meeting of the American Chemical Society, Miami Beach, Fla., Sept. 1978.
(11) Fisanick, W.; Mitchell, L. D.; Scott, J. A.; Vander Stouw, G. G. "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files", *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 73–84.
(12) Dunn, R. G.; Fisanick, W.; Zamora, A. "A Chemical Substructure Search System Based on Chemical Abstracts Index Nomenclature", *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 212–219.
(13) Kowalski, B. R.; Bender, C. F. "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test", *J. Am. Chem. Soc.* **1974**, *96*, 916–918.
(14) Chu, K. C.; Feldmann, R. J.; Shapiro, M. B.; Hazard, G. F.; Geran, R. I. "Pattern Recognition and Structure-Activity Relationship Studies. Computer-Assisted Prediction of Antitumor Activity in Structurally Diverse Drugs in an Experimental Mouse Brain Tumor System", *J. Med. Chem.* **1975**, *18*, 539–545.