Formal Grammar. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 106-112.
(9) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Nomenclature. 3. Syntax Analysis and Semantic Processing. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 112-118.
(10) Conrow, K. Computer Generation of Baeyer System Names of Saturated, Bridged, Bicyclic, Tricyclic, and Tetracyclic Hydrocarbons. *J. Chem. Doc.* **1966**, *6*, 206-212.
(11) Van Binnendyk, D.; MacKay, A. C. Computer-Assisted Generation of IUPAC Names of Polycyclic Bridged Ring Systems. *Can. J. Chem.* **1973**, *51*, 718-723.
(12) Vander Stouw, G. G.; Gustafson, C.; 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*, 213-218.
(13) Mockus, J.; Isenberg, A. C.; Vander Stouw, G. G. Algorithmic Generation of Chemical Abstracts Index Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 183-195.
(14) Meyer, D. E.; Gould, S. R. Microcomputer Generation of Chemical Nomenclature from Graphic Structure Input. *Am. Lab.* **1988**, *20* (11), 92-96.
(15) Meyer, D. E.; Warr, W. A.; Love, R. A. Chemical Structure Software for Personal Computers; ACS Professional Reference Book; American Chemical Society: Washington, DC, 1988.
(16) Wisniewski, J. L. Effective Text Compression with Simultaneous Digram and Trigram Encoding. *J. Inf. Sci.* **1987**, *13*, 159-164.
(17) Willet, P. A Review of Chemical Structure Retrieval Systems. *J. Chemom.* **1987**, *1*, 139-155.
(18) International Union of Pure and Applied Chemistry. Revision of the Extended Hantzsch-Widman System of Nomenclature for Heteromonocycles. *Pure Appl. Chem.* **1983**, *55* (2), 409-416.
(19) Tenenbaum, A. M.; Augenstein, M. J. *Data Structures Using Pascal.* Prentice-Hall: Englewood Cliffs, NJ, 1981; pp 252 and 318.

Topological Statistics on a Large Structural File

MICHEL PETITJEAN and JACQUES-ÉMILE DUBOIS*

Institut de Topologie et de Dynamique des Systèmes (ITODYS), associé au CNRS, Université de Paris VII, 1 rue Guy de la Brosse, 75005 Paris, France

Received March 7, 1990

Statistics based upon connection tables have been determined for a large structural file. Many distributions have unexpected local maxima and minima. Parity phenomena are observed in the distribution of hydrogen and carbon. An interpretation of even and odd distributions is proposed. Some of the compounds which represent topological extremes are shown.

INTRODUCTION

Many statistics in chemistry can be developed from chemical compounds that are, with their associated data, registered in databases, covering for example, spectroscopy, biological activity, thermodynamic properties, and so on. These parameters represent a source for numerous statistical investigations and research into correlations. When such investigations focus on the structural data for fully characterized compounds, very few analyses have been reported and they all depend upon Chemical Abstracts Service (CAS) for their source data.[1,2] When large files are involved, some aspects of our basic knowledge of chemical data depend largely upon these statistics. In this paper, we present original results derived from a CAS file[3] containing 3 424 428 compounds registered through July 1978.

Although statistics on cyclic and heterocyclic systems have been reported,[1,2] some complementary topological information in the structural data are reported here and provide valuable information for the chemist investigating large chemical datasets. Such information is useful for optimization of algorithms which require a statistical knowledge of topological data, such as atomic excentricities or concentric layers around a focus which can be used, for example, when applying the Cahn-Ingold-Prelog rules in computation of configuration. Some unexpected statistical results were obtained and these cannot be interpreted without the use of graph theory. Most of the statistical variables, therefore, will be derived from graph theory rather than from chemical considerations.

REPRESENTATION OF CHEMICAL COMPOUNDS

The expanded formulas of the compounds in the database are coded by means of a DARC-like[4] colored graph, but the statistical study is carried out without any preconceived idea concerning coding rules. The study is aimed essentially at extracting fundamental topological information from the file. The following colored graph terminology is used in the presentation of the results.

In a compound formula:

the graph nodes are the atoms

the graph edges are the chemical bonds

The atoms and bonds are both colored. Each atom assumes one of the 103 colors defined by the Mendeleev Table (the 103 atomic symbols), and each bond takes one of the following values: SI (simple), TA (tautomer), AR (aromatic), DO (double), or TR (triple).

In addition, the atoms are labeled with secondary chromatic information, and in this way, the complete description provided for each molecular structure includes

"Unusual" valency: arithmetic positive value between 0 and 99, a value of zero meaning the usual valency.

Charge: algebraic signed value between −9 and +9 associated with the delocalization flag; localized charge or not.

Isotope: arithmetic positive value giving the integer mass of the isotope. Zero means natural abundance.

Stereochemistry: may be included in the code, but is not registered in this part of the file.

0095-2338/90/1630-0332$02.50/0 © 1990 American Chemical Society

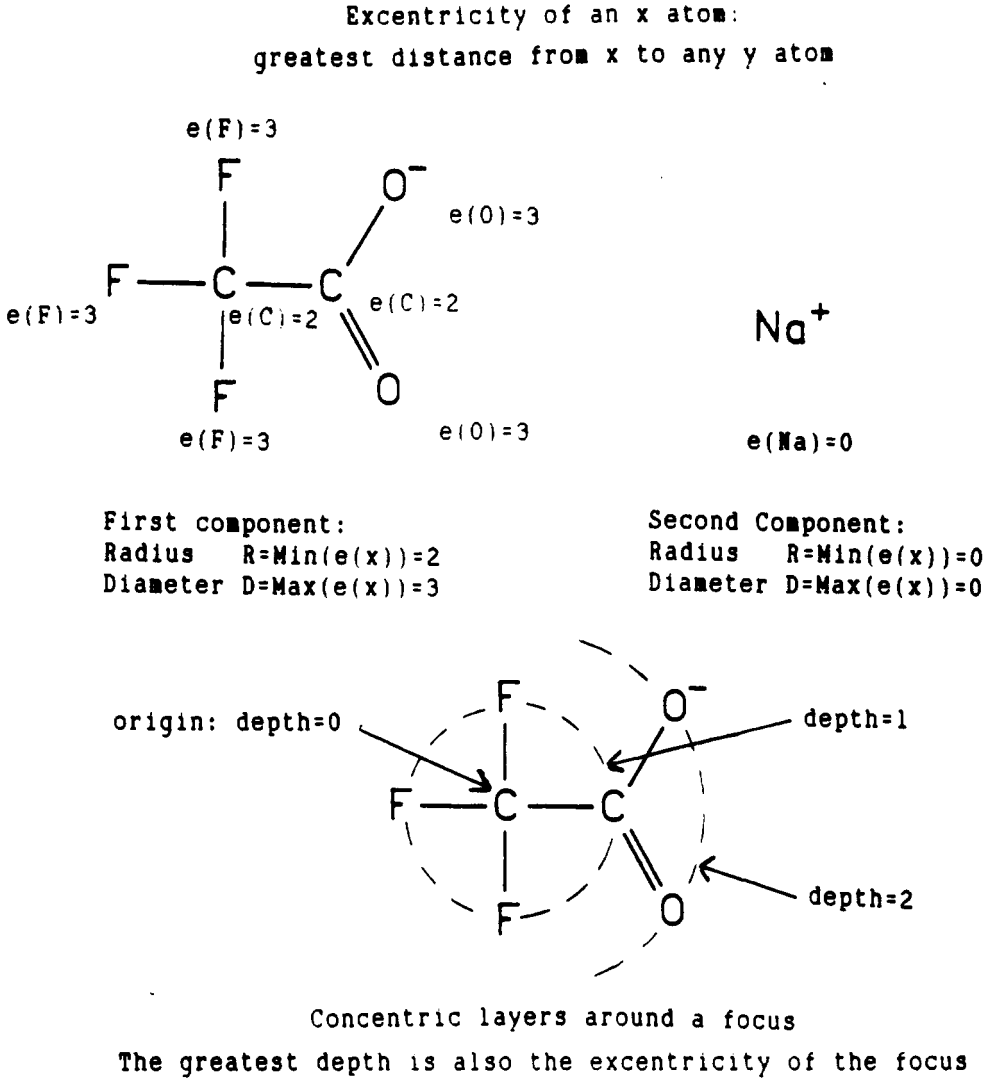

Figure 1. Different graph theory topological variables, exemplified by sodium trifluoroacetate.

Table I. Different Sets of Observations Considered in the File

type observations	total observations
chemical compounds	3 424 428
atoms	77 915 142
components	4 019 514
shortest path between atoms–pairs	2 225 906 690
concentric layers focused around atoms	898 037 816

A compound such as sodium fluoroacetate is comprised of *components*, in this case sodium ion and fluoroacetate ion. The components in the graph of a compound are labeled. Each component carries a ratio, which is the fractional multiplicity coefficient of the component, which, for the first registered component, is arbitrarily set at 1:1.

In addition, hydrogen atoms (except for 7957 hydrogens bearing secondary chromatisms) are implicit and are not recorded in the connection table. Thanks to the graph concept, this topological representation can express most of the expanded formulas correctly. Some variations on this concept (e.g., oriented multigraphs, recording of hydrogens, different sets of chromatisms) are possible and lead to modifications in the set of compounds that can be coded.[3] It was not felt to be necessary to code every compound in the file but only to code enough compounds to allow extraction of some of the robust statistical phenomena from the database.

STATISTICAL VARIABLES

The statistical variables can be divided into two categories: (a) Distributions dependent upon the particular coding of the graph used on a computer. These are essentially those that depend on the internal numbering of the atoms.

Table II. Number of Compounds with a Given Number of Components

components	compounds	components	compounds
1	2 863 557	7	79
2	533 323	8	57
3	22 471	9	5
4	4 128	10	5
5	557	11	3
6	242	12	1

(b) Distributions that depend only on the chemical information carried in each graph. The number of carbon atoms, for example, does not depend upon the local coding. It should also be noted that the interpretation of the distribution(s) depends partly on the set of nonrepresentable compounds. Every interpretation will be correct for our subfile of the CAS database, but some differences may be discerned when the full CAS file is considered. Only these latter distributions will be explored. In this way, numerous results of little value to chemists will be avoided.

If a set of primary distributions has been defined, it is always possible to build many new synthetic distributions, either univariate or multivariate, affording some complex results. In order to limit the quantity of combinatorial distributions, only some of the more important ones are given here. Among these important distributions, those concerning cycles and heterocycles have been published recently[1,2] and will not be reported

Table III. Number of Compounds with a Given Number of H Atoms

H	compds	H	compds	H	compds	H	compds	H	compds	H	compds
0	26667	52	10062	104	474	156	90	208	28	260	4
1	9414	53	3405	105	221	157	33	209	16	262	5
2	16229	54	8050	106	410	158	92	210	38	263	1
3	20177	55	2772	107	138	159	29	211	23	264	3
4	31957	56	6086	108	407	160	89	212	42	265	3
5	35985	57	2196	109	134	161	33	213	16	266	3
6	57808	58	4597	110	334	162	78	214	24	267	2
7	60601	59	1744	111	183	163	29	215	7	268	5
8	91650	60	3809	112	285	164	67	216	24	269	1
9	93766	61	1472	113	142	165	39	217	16	270	6
10	131064	62	3098	114	258	166	58	218	24	272	3
11	120266	63	1424	115	116	167	18	219	12	274	2
12	164729	64	2593	116	235	168	57	220	27	275	2
13	137056	65	1268	117	114	169	19	221	21	278	2
14	181755	66	2663	118	190	170	49	222	25	280	2
15	142664	67	1122	119	98	171	22	223	17	282	2
16	183715	68	2382	120	195	172	52	224	38	283	1
17	129539	69	1168	121	114	173	18	225	27	284	2
18	174683	70	2051	122	173	174	51	226	30	286	1
19	116053	71	1000	123	116	175	24	227	22	288	3
20	153599	72	1918	124	143	176	37	228	26	290	1
21	99849	73	1010	125	86	177	18	229	9	291	2
22	132270	74	1737	126	182	178	37	230	29	292	1
23	83084	75	1054	127	85	179	19	231	17	294	1
24	113151	76	1620	128	120	180	45	232	19	296	1
25	60046	77	735	129	79	181	9	233	16	297	2
26	94288	78	1635	130	110	182	43	234	25	298	1
27	56014	79	712	131	51	183	23	235	10	299	1
28	77274	80	1377	132	113	184	39	236	12	300	2
29	43855	81	701	133	61	185	17	237	15	302	1
30	67180	82	1241	134	108	186	27	238	20	304	2
31	34396	83	610	135	89	187	13	239	11	308	1
32	53559	84	1196	136	98	188	41	240	22	312	1
33	27018	85	588	137	64	189	15	241	9	318	1
34	44105	86	1131	138	117	190	32	242	7	320	2
35	20459	87	506	139	52	191	18	243	9	323	1
36	35731	88	877	140	102	192	49	244	11	326	1
37	15839	89	460	141	53	193	15	245	7	330	2
38	27129	90	948	142	107	194	31	246	10	342	2
39	13070	91	420	143	44	195	14	247	4	346	1
40	21244	92	681	144	112	196	40	248	13	348	1
41	10049	93	458	145	54	197	12	249	5	350	1
42	18383	94	666	146	110	198	29	250	8	352	1
43	8659	95	335	147	55	199	14	251	7	358	1
44	15761	96	559	148	140	200	29	252	14	360	2
45	7796	97	311	149	67	201	17	253	6	368	1
46	14764	98	516	150	90	202	26	254	3	383	1
47	6334	99	280	151	57	203	11	255	4	384	1
48	13813	100	538	152	77	204	32	256	2	418	1
49	5402	101	191	153	46	205	10	257	2	450	1
50	12375	102	477	154	85	206	31	258	7		
51	4666	103	211	155	49	207	15	259	2		

Table IV. Number of Compounds with a Given Number of D or T Atoms

D or T	D compds	T compds
0	3423017	3424316
1	575	100
2	525	5
3	155	4
4	90	0
5	28	0
6	23	0
7	7	1
8	3	1
9	1	0
10	2	0
12	1	0
17	1	1

Table V. Number of Atoms with a Given Connection Degree

connection degree (no. of neighbors)	atoms	connection degree (no. of neighbors)	atoms
0	403027	8	629
1	16814229	9	227
2	37072401	10	3281
3	21202169	11	45
4	2357230	12	89
5	49783	13	5
6	11825	14	4
7	198		

here. The aim of this paper is to complement the published results and, hopefully, provide a different insight into the chemical information that is in the file. Some brief comparisons between our statistics on the 1978 file and the different statistical results derived by CAS from the 1974, 1979, and 1987 files show only minor variations. An exception concerns the number of compounds containing either deuterium or tritium. Such compounds occur less frequently in the 1978 file.

The file can be considered as a set of five different types of observation (see Table I). Each of these five types of observation is a possible topological unit, and they will be examined in turn. In order to avoid confusion, some classical

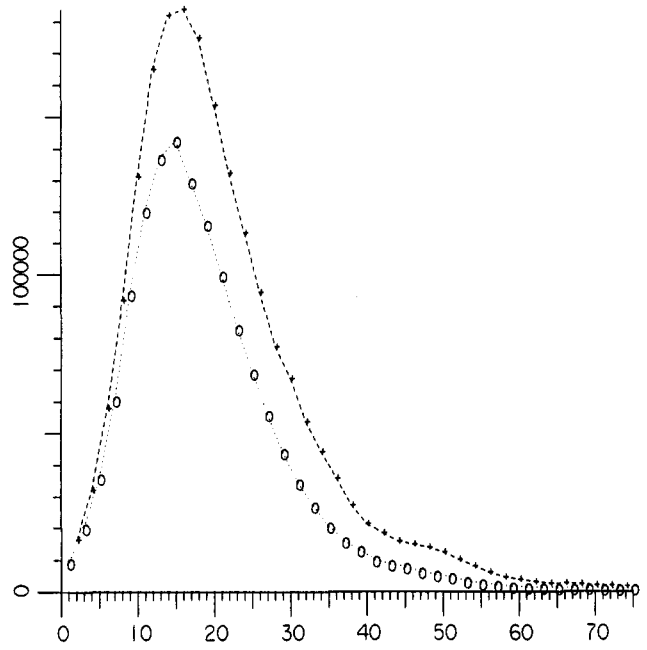

Figure 2. Hydrogen distribution, concerning compounds (Table III). The upper curve corresponds to even values, with an absolute maximum for 16 hydrogens (183 715 compounds). Only the range 1–75 is displayed: highest values represent less than 0.9% of the hydrogen-containing compounds.

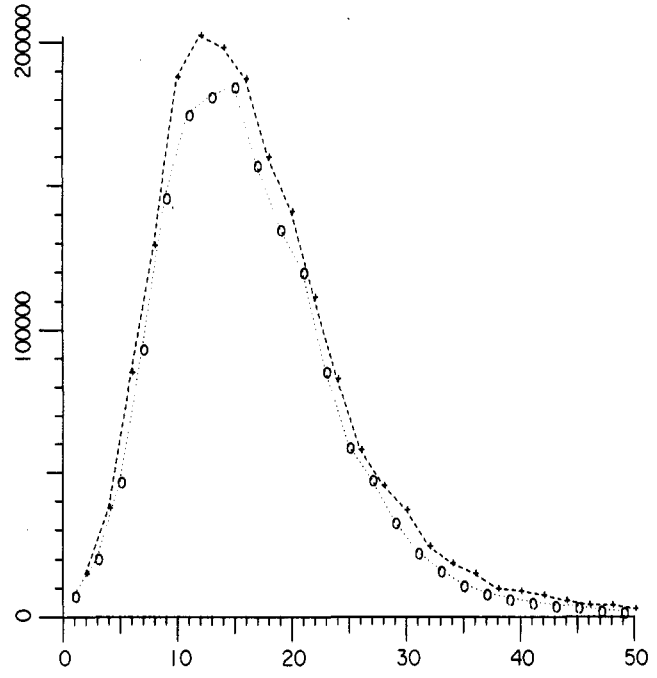

Figure 3. Carbon distribution, concerning compounds [data from (ref 3)]. The upper curve corresponds to even values, with an absolute maximum for 12 carbons (202 276 compounds). Only the range 1–50 is displayed: highest values represent less than 1% of the carbon-containing compounds.

terms for nonoriented graphs are defined here following Berge.[5] These terms are also illustrated in Figure 1.

(1) The topological "distance" between two nodes (atoms) x and y is the minimum number of edges (chemical bonds) between the two nodes. Trifluoroacetic acid, for example, has a distance d(F,O) of 3 between any of the fluorines and any of the oxygens. This distance is a mathematical distance which satisfies the three axioms:

$$d(x,y) \geq 0 \text{ and } d(x,y) = 0 <=> x = y$$

$$d(x,y) = d(y,x)$$

$$d(x,y) \geq d(x,z) + d(z,y) \quad \text{(triangular inequality)}$$

Table VI. Number of Atoms with a Given Excentricity E

E	atoms	E	atoms	E	atoms
0	403 027	50	22 132	100	285
1	109 369	51	19 725	101	178
2	468 417	52	20 115	102	196
3	1 133 319	53	20 433	103	162
4	2 964 236	54	17 315	104	159
5	5 335 970	55	17 120	105	142
6	7 435 105	56	16 553	106	117
7	8 412 779	57	15 127	107	99
8	8 634 264	58	15 461	108	138
9	8 012 310	59	14 155	109	80
10	6 820 316	60	12 808	110	84
11	5 501 553	61	11 536	111	49
12	4 350 907	62	10 948	112	58
13	3 398 567	63	9 837	113	48
14	2 676 495	64	9 540	114	59
15	2 055 809	65	9 300	115	52
16	1 629 633	66	8 152	116	54
17	1 308 694	67	8 138	117	49
18	1 045 464	68	7 698	118	56
19	849 553	69	6 090	119	45
20	694 321	70	6 480	120	48
21	577 229	71	6 001	121	44
22	491 954	72	5 985	122	50
23	415 228	73	5 912	123	40
24	354 948	74	5 254	124	41
25	307 910	75	4 051	125	36
26	263 759	76	3 917	126	34
27	229 265	77	3 761	127	32
28	201 747	78	3 236	128	22
29	178 890	79	2 878	129	22
30	152 623	80	2 733	130	20
31	136 269	81	2 685	131	14
32	118 865	82	2 800	132	14
33	103 773	83	2 341	133	16
34	95 490	84	2 464	134	14
35	87 563	85	2 195	135	12
36	77 472	86	2 097	136	12
37	69 122	87	2 083	137	12
38	63 174	88	1 930	138	10
39	54 804	89	1 716	139	8
40	51 488	90	1 473	140	6
41	46 829	91	1 318	141	8
42	39 861	92	846	142	6
43	36 231	93	729	143	6
44	33 429	94	707	144	6
45	29 425	95	527	145	12
46	29 232	96	485	146	4
47	27 972	97	397	147	0
48	24 171	98	396	148	0
49	23 825	99	247	149	0

When there is no path between the two nodes, the distance $d(x,y)$ is considered to be infinite. For example, sodium trifluoroacetate has no path between the sodium and the other atoms, so d(Na,C), d(Na,F), and d(Na,O) are all infinite. Bond multiplicities do not affect the values of the distances.

(2) A finite value for the distance $d(x,y)$ defines a relationship between x and y which is reflexive, symmetrical, and transitive. There are thus equivalence classes: the subgraph containing all the nodes of a class is called a *component*. Sodium trifluoroacetate, as noted, has two components: the trifluoroacetate ion and the sodium ion.

(3) In a given component, the *excentricity* $e(x)$ of the node x is the maximum value of $d(x,y)$ taken in the set of all y nodes in the component. It is also the maximum number of concentric layers focused on x (see Figure 1). In trifluoroacetate ion, e(F) = e(O) = 3 for the three fluorines and the two oxygens, and e(C) = 2 for the two carbons. Every isolated atom, such as the sodium ion, has zero excentricity: e(Na) = 0.

(4) The radius R of a component is the minimum excentricity $e(x)$, taken from the set of all the x nodes. The node for which $e(x) = R$ is called the *centroid* of the component,

Table VII. Number of Components with a Given Number of Atoms

atoms	compns	atoms	compns	atoms	compns	atoms	compns	atoms	compns
1	403027	52	4860	103	255	154	41	205	19
2	9199	53	3670	104	269	155	53	206	22
3	13818	54	4299	105	251	156	35	207	30
4	30997	55	3174	106	215	157	58	208	65
5	58200	56	3648	107	178	158	43	209	24
6	38069	57	2870	108	212	159	46	210	26
7	38526	58	3284	109	195	160	41	211	25
8	58142	59	2514	110	165	161	44	212	25
9	63719	60	2947	111	183	162	39	213	18
10	88998	61	2285	112	166	163	33	214	16
11	109012	62	2341	113	126	164	27	215	28
12	128170	63	1977	114	151	165	36	216	26
13	143647	64	2088	115	195	166	36	217	27
14	154670	65	1592	116	177	167	37	218	18
15	158922	66	1917	117	155	168	26	219	25
16	193884	67	1433	118	143	169	36	220	20
17	169810	68	1530	119	148	170	40	221	18
18	176002	69	1192	120	140	171	32	222	16
19	167672	70	1280	121	107	172	31	223	12
20	170321	71	1054	122	124	173	43	224	12
21	159130	72	1326	123	111	174	29	225	17
22	156370	73	961	124	111	175	32	226	15
23	138869	74	1134	125	92	176	32	227	15
24	134663	75	905	126	109	177	39	228	17
25	116668	76	1037	127	92	178	34	229	19
26	108962	77	1084	128	95	179	32	230	14
27	90361	78	976	129	93	180	34	231	16
28	86570	79	690	130	70	181	23	232	12
29	73404	80	767	131	71	182	25	233	12
30	68408	81	698	132	80	183	20	234	24
31	56955	82	701	133	85	184	19	235	19
32	53318	83	591	134	83	185	25	236	11
33	42737	84	725	135	84	186	26	237	24
34	39971	85	630	136	65	187	32	238	25
35	31240	86	626	137	91	188	23	239	38
36	30003	87	469	138	61	189	25	240	19
37	22993	88	542	139	57	190	28	241	25
38	22459	89	428	140	51	191	36	242	21
39	17453	90	475	141	71	192	19	243	21
40	17600	91	419	142	51	193	17	244	19
41	13751	92	453	143	42	194	15	245	16
42	14296	93	437	144	60	195	22	246	19
43	10970	94	422	145	72	196	27	247	5
44	11894	95	365	146	65	197	26	248	17
45	8848	96	405	147	34	198	14	249	8
46	9033	97	286	148	52	199	27	250	26
47	6826	98	313	149	65	200	20	251	19
48	7576	99	263	150	64	201	60	252	17
49	5473	100	304	151	50	202	26	253	17
50	5943	101	274	152	45	203	23	254	0
51	4626	102	279	153	67	204	19	255	0

but it is often not unique. Each centroid is a focus minimizing the number of concentric layers around it. The two carbons of trifluoroacetate ion are both centroids, and the radius of this ion is $R = 2$. The radius of an isolated node, such as the sodium ion, is always $R = 0$.

(5) The diameter D is a maximum of $e(x)$, taken from the set of all the x nodes. The nodes for which $e(x) = D$ are called *extremal nodes*. By applying triangular inequality to the centroid and the extremal nodes, it may be shown that D varies between R and $2R$, depending on the component. Each extremal atom is a focus maximizing the number of concentric layers around it. Trifluoroacetate ion for example, has a diameter of $D = 3$. The diameter of an isolated atom, like the sodium ion, is always $D = R = 0$. For acyclic components with an even diameter, the centroid is unique and $D = 2R$. For acyclic components with an odd diameter, $D = 2R - 1$, and there are always two centroids $x1$ and $x2$ with $d(x1,x2) = 1$.

For most hydrocarbons and their derivatives, chemical nomenclature is closely related to the value of D, which leads to the alkane series (containing $D + 1$ carbons) describing the component. When $D = 0$ the compound will be a methane derivative, $D = 1$ is an ethane derivative, and so on.

DISTRIBUTIONS IN COMPOUNDS

A chemical compound is the natural unit usually considered by the chemist. The following statistics are derived from a database of 3424428 compounds (except for the carbon distribution, which is based on the 3387025 carbon-containing compounds in the file). The bond distribution in the file is as follows: simple, 56.0%; tautomer, 5.4%; aromatic, 30.9%; double, 7.4%; triple, 0.3%. The bond chromatism, in particular the choice between aromatic and tautomeric, is computed by the coding algorithm which uses a set of rules that may not be optimal for all complex systems. The bond chromatism depends on this set of rules and thus on the coding algorithm. The charge distribution, which is sometimes difficult to define in delocalized systems, is in the same situation, as is the valency distribution. The concept of "usual valency" is unclear when applied to metallic elements which have numerous oxidation numbers and which can form chelates with different numbers of neighbors.

Table VIII. Number of Components with a Given Number of C Atoms

C	compns	C	compns	C	compns	C	compns	C	compns
0	483571	45	3441	90	192	135	31	180	5
1	21838	46	3893	91	116	136	66	181	2
2	40151	47	2313	92	120	137	50	182	1
3	25917	48	3933	93	113	138	27	183	3
4	61202	49	1867	94	108	139	29	184	0
5	52961	50	2683	95	116	140	33	185	1
6	133323	51	1678	96	170	141	37	186	0
7	108259	52	2072	97	100	142	63	187	1
8	139671	53	1263	98	146	143	23	188	1
9	152953	54	2096	99	119	144	36	189	3
10	198105	55	1312	100	126	145	36	190	3
11	178233	56	1588	101	97	146	32	191	1
12	211654	57	1116	102	95	147	28	192	0
13	184800	58	1214	103	56	148	38	193	0
14	200481	59	747	104	79	149	19	194	0
15	185439	60	1346	105	66	150	43	195	0
16	189068	61	680	106	83	151	44	196	2
17	157393	62	996	107	56	152	21	197	0
18	160877	63	705	108	95	153	22	198	1
19	134878	64	865	109	61	154	33	199	1
20	140627	65	583	110	73	155	21	200	1
21	119184	66	743	111	61	156	17	201	0
22	108950	67	375	112	48	157	15	202	0
23	83453	68	659	113	46	158	19	203	0
24	79712	69	419	114	61	159	24	204	0
25	55998	70	540	115	45	160	16	205	0
26	54535	71	311	116	47	161	7	206	1
27	45319	72	692	117	51	162	9	207	0
28	43128	73	278	118	43	163	10	208	1
29	30869	74	306	119	29	164	12	209	0
30	34897	75	277	120	66	165	5	210	0
31	20818	76	438	121	38	166	4	211	0
32	22814	77	238	122	47	167	3	212	0
33	15114	78	339	123	35	168	11	213	0
34	17259	79	193	124	34	169	9	214	0
35	10437	80	329	125	41	170	7	215	0
36	14417	81	232	126	36	171	3	216	0
37	7848	82	247	127	26	172	2	217	0
38	9229	83	132	128	29	173	8	218	0
39	6001	84	270	129	34	174	2	219	0
40	8568	85	137	130	46	175	4	220	1
41	4789	86	179	131	46	176	1	221	0
42	7317	87	132	132	50	177	6	222	0
43	3819	88	210	133	23	178	3	223	0
44	5271	89	131	134	42	179	0	224	0

	mean	SD	local maxima	local minima
no. of atoms	22.753	16.626	18	
no. of bonds	23.998	13.776	0;8	2;10
simple bonds	13.435	9.760	16	2
tautomer bonds	1.290	2.940	0;2;4	1;3;5
aromatic bonds	7.417	7.497	0;6;12;18;24;30 ...(6*i)...	
double bonds	1.784	1.866	0	
triple bonds	0.072	0.347	0	
heteroatoms	5.953	4.822	4 (sharp)	
"unusual" valencies	0.390	0.987	0	
no. of charges	0.109	0.530	0	
no. of "isotopes"	0.006	0.107	0	
no. of characters in mol formula	10.157	2.797	8;10	9
hydrogens (mol formula)	20.706	14.013	14;16	15
carbons	16.985	9.561	12;14	13

The complete distributions for the number of compounds with a given number of components are given in Table II. The distribution of compounds with a given number of hydrogen atoms is given in Table III and Figure 2 and for a given number of deuterium and tritium atoms in Table IV. Data concerning carbon distribution (the numbers of compounds having specific numbers of carbon atoms) was published previously[3] and is displayed in Figure 3.

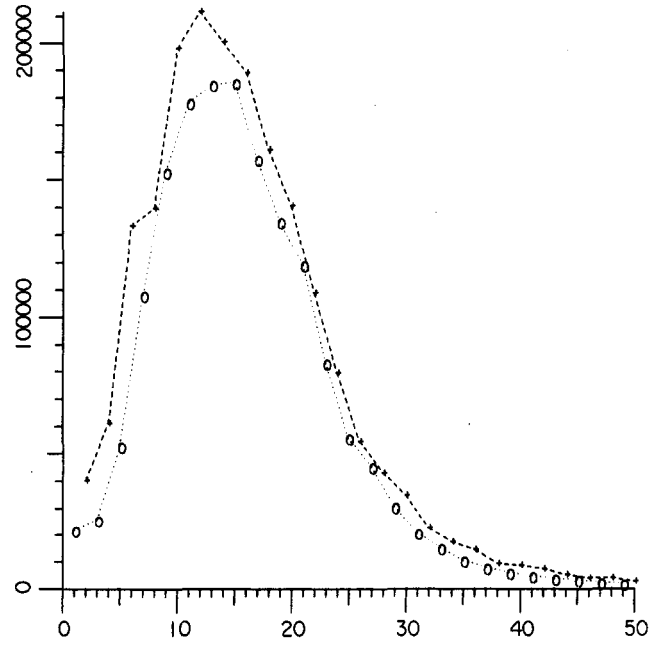

Figure 4. Carbon distribution, concerning components (Table VIII). The upper curve corresponds to even values, with an absolute maximum for 12 carbons (211 654 components). Only the range 1–50 is displayed: highest values represent less than 0.9% of the carbon-containing components.

Table IX. Number of Components with a Given Number of Bonds

bonds	compns	bonds	compns	bonds	compns	bonds	compns	bonds	compns	bonds	compns
0	403 027	48	10 642	96	511	144	62	192	28	240	16
1	9 199	49	8 346	97	421	145	54	193	30	241	17
2	13 538	50	7 950	98	404	146	58	194	36	242	32
3	30 673	51	7 007	99	381	147	59	195	22	243	21
4	56 357	52	6 788	100	353	148	62	196	28	244	16
5	31 340	53	5 537	101	333	149	59	197	18	245	21
6	33 565	54	5 322	102	341	150	64	198	25	246	38
7	47 645	55	4 610	103	334	151	47	199	22	247	24
8	48 743	56	4 515	104	339	152	71	200	31	248	23
9	65 612	57	3 923	105	305	153	56	201	27	249	15
10	74 412	58	3 596	106	252	154	47	202	22	250	13
11	97 394	59	3 198	107	258	155	45	203	27	251	18
12	115 798	60	3 523	108	274	156	82	204	25	252	18
13	119 796	61	2 944	109	229	157	36	205	26	253	19
14	130 453	62	2 846	110	238	158	58	206	35	254	10
15	139 455	63	2 813	111	234	159	44	207	21	255	21
16	170 998	64	2 669	112	244	160	74	208	46	256	7
17	145 322	65	2 292	113	194	161	42	209	15	257	6
18	149 651	66	2 322	114	188	162	48	210	23	258	11
19	151 243	67	2 027	115	200	163	36	211	14	259	12
20	150 199	68	1 975	116	182	164	40	212	22	260	19
21	145 982	69	1 933	117	186	165	27	213	25	261	10
22	139 803	70	1 608	118	180	166	35	214	14	262	6
23	133 189	71	1 551	119	169	167	34	215	28	263	4
24	129 867	72	1 849	120	186	168	52	216	63	264	4
25	122 824	73	1 265	121	136	169	24	217	33	265	7
26	113 002	74	1 157	122	145	170	35	218	22	266	5
27	105 423	75	1 224	123	176	171	31	219	21	267	9
28	94 946	76	1 114	124	142	172	48	220	26	268	4
29	84 592	77	1 084	125	118	173	30	221	12	269	5
30	78 780	78	1 129	126	123	174	48	222	26	270	2
31	69 307	79	935	127	107	175	31	223	30	271	2
32	64 629	80	962	128	112	176	34	224	18	272	5
33	57 075	81	844	129	103	177	31	225	19	273	5
34	51 943	82	751	130	98	178	35	226	20	274	3
35	46 209	83	775	131	108	179	19	227	16	275	11
36	42 117	84	1 059	132	130	180	46	228	23	276	3
37	34 456	85	657	133	88	181	28	229	9	277	6
38	30 929	86	625	134	106	182	30	230	26	278	8
39	27 140	87	620	135	96	183	18	231	14	279	3
40	24 315	88	651	136	96	184	36	232	25	280	3
41	20 089	89	566	137	64	185	26	233	21	281	1
42	19 234	90	687	138	105	186	19	234	10	282	6
43	15 645	91	555	139	72	187	44	235	9	288	1
44	14 690	92	604	140	82	188	30	236	17		
45	13 149	93	466	141	77	189	22	237	16		
46	11 951	94	470	142	90	190	22	238	23		
47	10 424	95	427	143	57	191	20	239	6		

Table X. Number of Components with a Given Number of Cycles

cycles	compns	cycles	compns	cycles	compns
0	931 825	21	423	42	4
1	785 193	22	112	43	0
2	874 538	23	99	44	0
3	667 176	24	92	45	0
4	426 887	25	71	46	0
5	184 585	26	51	47	0
6	80 281	27	50	48	4
7	29 200	28	33	49	0
8	15 145	29	47	50	0
9	7 331	30	52	51	2
10	5 320	31	12	52	0
11	2 567	32	17	53	2
12	2 323	33	9	54	0
13	1 022	34	1	55	0
14	634	35	6	56	0
15	605	36	14	57	17
16	413	37	1	58	2
17	244	38	127	59	3
18	205	39	52	60	3
19	1 756	40	46	61	0
20	908	41	4	62	0

Table XI. Number of Components with a Given Minimum Nodal Connection Degree (Range 0–5) and a Maximum Nodal Connection Degree (Range 0–14)

	0	1	2	3	4	5	total
0	403 027						403 027
1		9 199					9 199
2		38 419	11 248				49 667
3		2 047 980	109 229	38			2 157 247
4		1 354 393	25 338	5	15		1 379 751
5		8 258	536	28	124	31	8 977
6		7 137	397	1	14	6	7 555
7		95	17	0	0	0	112
8		391	127	1	0	0	519
9		125	31	1	0	0	157
10		2 879	258	19	4	0	3 160
11		33	7	3	2	0	45
12		78	10	1	0	0	89
13		5	0	0	0	0	5
14		4	0	0	0	0	4
total	403 027	3 468 996	147 198	97	159	37	4 019 514

PARITY PHENOMENA

The distributions of each of the 103 elements were calculated. It was observed that most of them decrease rapidly and monotonically. Some exceptions are boron with local maxima of 3, 6, and 12; oxygen, with a local maximum of 2; and fluorine, with local maxima of 3, 6, and 12; but the major exceptions, as mentioned previously,[3] are hydrogen and carbon. The frequency curve can be divided into two parts, one related

Table XII. Number of Components with a Given Radius

R	compns	R	compns	R	compns
0	403 027	25	745	50	31
1	99 542	26	523	51	9
2	101 434	27	447	52	9
3	403 971	28	513	53	2
4	671 409	29	315	54	11
5	737 589	30	362	55	10
6	604 704	31	263	56	4
7	377 290	32	277	57	1
8	227 990	33	201	58	5
9	144 267	34	230	59	3
10	82 197	35	161	60	4
11	48 431	36	115	61	1
12	31 285	37	187	62	0
13	21 507	38	205	63	6
14	14 520	39	130	64	2
15	10 997	40	136	65	2
16	8 265	41	57	66	0
17	5 821	42	75	67	3
18	5 799	43	59	68	3
19	3 887	44	78	69	1
20	3 826	45	63	70	1
21	2 314	46	157	71	1
22	1 788	47	51	72	0
23	1 211	48	37	73	3
24	896	49	48	74	0

Table XIII. Number of Components with a Given Diameter

D	compns	D	compns	D	compns
0	403 027	50	329	100	21
1	9 483	51	269	101	1
2	92 319	52	245	102	6
3	48 924	53	246	103	5
4	92 432	54	173	104	1
5	168 808	55	281	105	2
6	265 915	56	198	106	0
7	280 176	57	157	107	0
8	346 557	58	145	108	11
9	359 507	59	190	109	4
10	373 907	60	165	110	4
11	316 658	61	160	111	0
12	276 997	62	92	112	0
13	204 392	63	104	113	0
14	169 564	64	174	114	1
15	128 014	65	102	115	0
16	98 343	66	97	116	0
17	81 188	67	138	117	0
18	60 101	68	89	118	1
19	46 499	69	60	119	1
20	35 240	70	96	120	1
21	26 878	71	37	121	0
22	21 044	72	68	122	1
23	17 062	73	108	123	0
24	12 733	74	77	124	0
25	12 627	75	152	125	1
26	8 658	76	49	126	5
27	8 146	77	82	127	2
28	5 895	78	46	128	0
29	6 113	79	73	129	1
30	4 515	80	49	130	1
31	4 591	81	22	131	0
32	3 357	82	32	132	0
33	3 087	83	28	133	0
34	2 375	84	45	134	1
35	2 916	85	28	135	1
36	2 604	86	28	136	0
37	2 142	87	32	137	0
38	1 642	88	40	138	1
39	2 071	89	20	139	1
40	1 611	90	36	140	0
41	1 212	91	126	141	0
42	1 052	92	26	142	1
43	1 003	93	11	143	0
44	745	94	37	144	0
45	637	95	14	145	0
46	572	96	20	146	2
47	444	97	20	147	0
48	397	98	26	148	0
49	406	99	9	149	0

Table XIV. Number of Components with a Given Number of Centroids

centroids	compns	centroids	compns	centroids	compns
1	1 970 024	51	2	101	0
2	1 370 650	52	12	102	1
3	373 873	53	2	103	0
4	210 908	54	10	104	2
5	51 341	55	1	105	0
6	19 556	56	12	106	1
7	5 203	57	2	107	0
8	4 204	58	4	108	2
9	2 473	59	0	109	0
10	2 063	60	11	110	1
11	1 195	61	0	111	0
12	1 466	62	0	112	1
13	653	63	5	113	1
14	1 020	64	8	114	0
15	551	65	2	115	0
16	692	66	2	116	0
17	257	67	0	117	0
18	512	68	4	118	1
19	144	69	2	119	0
20	412	70	4	120	1
21	159	71	0	121	0
22	272	72	13	122	1
23	98	73	0	123	0
24	321	74	1	124	1
25	76	75	0	125	0
26	178	76	1	126	1
27	74	77	1	127	0
28	140	78	2	128	0
29	48	79	0	129	0
30	156	80	6	130	0
31	21	81	8	131	0
32	143	82	1	132	0
33	26	83	1	133	0
34	50	84	4	134	1
35	17	85	1	135	0
36	104	86	1	136	1
37	6	87	0	137	0
38	31	88	4	138	0
39	7	89	2	139	0
40	51	90	3	140	1
41	2	91	0	141	0
42	42	92	5	142	0
43	3	93	0	143	0
44	36	94	1	144	0
45	11	95	0	145	0
46	18	96	3	146	0
47	3	97	0	147	0
48	34	98	3	148	0
49	8	99	0	149	0
50	19	100	2	150	0

to even values and the other to odd values (see Figures 2 and 3). The two frequency subcurves have similar shapes, with relative weights $W_e + W_o = 100\%$:

Hydrogens: $W_e - W_o = 17.57\%$ (26 667 compounds with no H are not included)

Carbons: $W_e - W_o = 6.04\%$ (37 403 compounds with no C are not included)

A possible explanation of this bias toward even values can be found in graph theory. The number of odd-connected nodes in a graph is always even.[5] Since the connection degrees of the nodes are the valencies of the elements (a double bond is represented with two edges, so very few carbons will not be tetraconnected), then the set of odd-connected atoms will contain a large number (70 908 564) of hydrogen atoms, most of which are monoconnected. At a lower frequency 10 568 323 nitrogen atoms, most triconnected, and 1 110 863 chlorine atoms, most monoconnected, can be found. If monovalent hydrogen atoms were the only monoconnected atoms, one should observe 100% of even-hydrogen-containing compounds, but since monovalent hydrogen is merely the most abundant odd-connected element, one observes only 58.8% of even-hydrogen-containing compounds.

Table XV. Number of Components with a Given Number of Extremal Atoms

ext atoms	compns	ext atoms	compns	ext atoms	compns
1	403 027	51	2	101	0
2	1 204 847	52	8	102	1
3	1 035 109	53	2	103	0
4	809 145	54	6	104	1
5	266 572	55	1	105	0
6	167 416	56	11	106	0
7	65 846	57	0	107	0
8	33 472	58	3	108	1
9	13 352	59	0	109	0
10	5 784	60	8	110	1
11	1 328	61	0	111	0
12	6 145	62	0	112	1
13	604	63	3	113	0
14	1 642	64	6	114	0
15	702	65	2	115	0
16	645	66	1	116	0
17	170	67	0	117	0
18	1 003	68	3	118	1
19	110	69	0	119	0
20	341	70	4	120	1
21	271	71	0	121	0
22	144	72	9	122	0
23	43	73	0	123	0
24	322	74	0	124	0
25	41	75	0	125	0
26	99	76	0	126	1
27	71	77	1	127	0
28	493	78	3	128	0
29	35	79	0	129	0
30	130	80	3	130	0
31	12	81	8	131	0
32	111	82	0	132	0
33	24	83	1	133	0
34	36	84	2	134	1
35	16	85	1	135	0
36	96	86	1	136	1
37	5	87	0	137	0
38	20	88	4	138	0
39	11	89	1	139	0
40	28	90	1	140	0
41	2	91	0	141	0
42	42	92	0	142	0
43	0	93	0	143	0
44	28	94	1	144	0
45	11	95	0	145	0
46	12	96	2	146	0
47	1	97	0	147	0
48	28	98	0	148	0
49	8	99	0	149	0
50	6	100	0	150	0

A similar explanation may hold for the carbon distribution. If the connection degrees of the nodes are the numbers of neighbors of the elements (double bonds are represented here with only one edge, so carbons may be either even- or odd-connected), then the set of odd-connected atoms contains a large number of triconnected carbons—57 528 231 carbons in the file, 57% of which are sp^2.[6] This is supported by the high number of benzenoid compounds; the benzene ring is by far the most common ring system in the database, about 20 times more common than the next most abundant ring (pyridine).[2] Monoconnected or pentaconnected carbons are very rare. There is also a large set of monoconnected hydrogens and, at a much lower occurrence frequency, odd-connected oxygens (5 820 786 in the database, some of them, such as alcohols and other hydroxy compounds, being even-connected). There are also odd-connected nitrogens and so on. If sp^2 (triconnected) carbons and monovalent hydrogens were the only odd-connected elements, one would observe 100% of even- (carbon and hydrogen) containing compounds and thus 100% of even-carbon-containing compounds because, as shown previously, the number of hydrogens would also be even. Since sp^2 (triconnected) carbons and monovalent hydrogens are the most abundant odd-connected elements, one observes only 53.0% of even-carbon-containing compounds. The parity phenomenon is also observed in connection with the atom distribution, i.e., the number of compounds with a given number of atoms, excluding hydrogen. The difference $W_e - W_o = 2.0\%$ is weak and may be related to the parity of the carbon distribution.

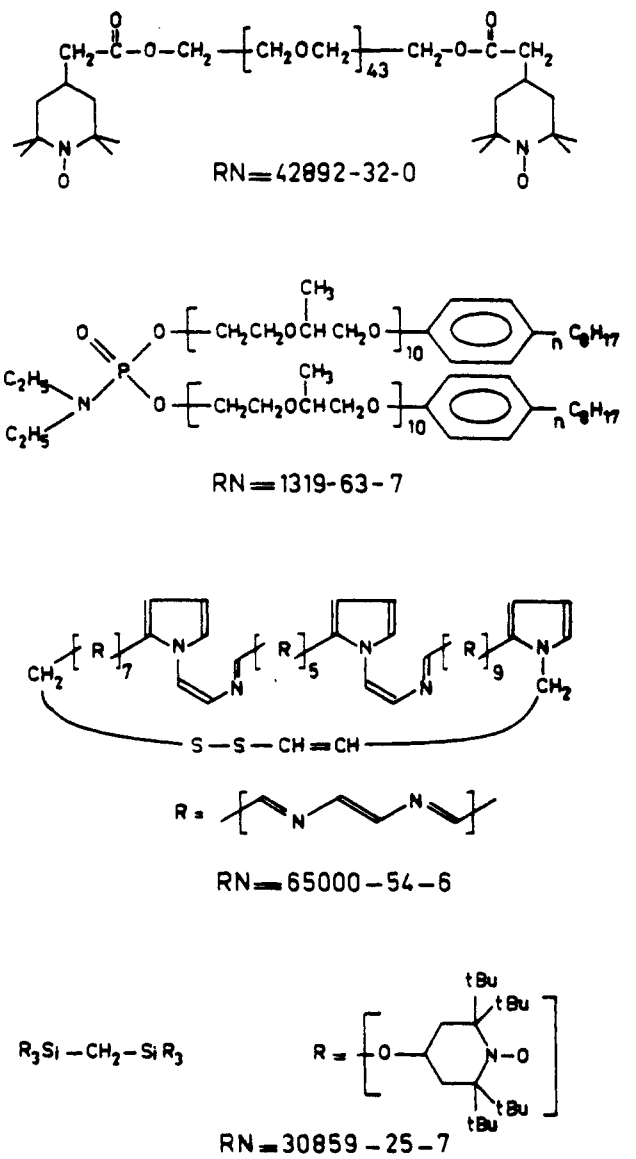

Figure 5. Some topologically extremal compounds in the file. Polyethers RN = 1319-63-7 and RN = 42892-32-0: diameter D = 146 and radius R = 73. Polyimine RN = 65000-54-6: R = 73, with 140 centroids. When focused on the methylene group, the deepest concentric layer of the silane RN = 30859-25-7 has 78 atoms.

DISTRIBUTIONS OF ATOMS

The second type of observation to be investigated is the set of all atoms. It may be recalled that hydrogens (except for the 7957 special hydrogen atoms mentioned previously) were not recorded and that multiple bonds were recorded as simple edges. The following statistics refer to the 77 915 142 atoms:

	mean	SD	local maxima	local minima
connection degree (no. of neighbors)	2.109	0.793	2;10	7
"unusual" valency	0.094	0.994	5	26
algebraic charge	0.172	2.526	+1	
isoptope	0.016	1.506	14	
excentricity	10.526	6.829	0;8	1

Table XVI. Number of Paths with a Given Length

length	paths	length	paths	length	paths
0	77915142	49	932612	98	5532
1	164357002	50	904376	99	4542
2	229244876	51	822726	100	4428
3	237692512	52	764290	101	3692
4	221359272	53	726886	102	3426
5	202636784	54	666188	103	2866
6	179934712	55	611690	104	2758
7	152706078	56	583028	105	2358
8	127099050	57	525988	106	2184
9	102337698	58	482148	107	1840
10	82488910	59	452832	108	1866
11	66382512	60	406912	109	1548
12	53557820	61	371236	110	1560
13	43501662	62	353212	111	1332
14	36087700	63	317674	112	1304
15	29671486	64	292086	113	1146
16	25029242	65	276096	114	1138
17	21479812	66	247612	115	1008
18	18251854	67	221624	116	972
19	15719910	68	208694	117	872
20	13859090	69	187798	118	840
21	12031596	70	176378	119	726
22	10584384	71	161236	120	694
23	9451346	72	144684	121	614
24	8251666	73	128382	122	578
25	7330208	74	119352	123	516
26	6692854	75	104168	124	456
27	5926080	76	95674	125	392
28	5320062	77	87854	126	348
29	4847706	78	78998	127	306
30	4314768	79	71760	128	258
31	3890326	80	68252	129	230
32	3634336	81	60444	130	204
33	3274204	82	55406	131	180
34	3002064	83	50674	132	166
35	2787416	84	45004	133	152
36	2498736	85	39322	134	134
37	2278066	86	35996	135	120
38	2168938	87	29642	136	110
39	1956388	88	25780	137	94
40	1786716	89	20948	138	84
41	1675338	90	17510	139	74
42	1521776	91	14002	140	80
43	1405354	92	12386	141	60
44	1362388	93	10258	142	64
45	1243010	94	9138	143	56
46	1160560	95	7750	144	64
47	1099932	96	7022	145	24
48	1003772	97	5848	146	4

Table XVII. Number of Layers with a Given Depth

depth	layers	depth	layers	depth	layers
0	77915142	49	376311	98	3302
1	77512115	50	352486	99	2906
2	77402746	51	330354	100	2659
3	76934329	52	310629	101	2374
4	75801010	53	290514	102	2196
5	72836774	54	270081	103	2000
6	67500804	55	252766	104	1838
7	60065699	56	235646	105	1679
8	51652920	57	219093	106	1537
9	43018656	58	203966	107	1420
10	35006346	59	188505	108	1321
11	28186030	60	174350	109	1183
12	22684477	61	161542	110	1103
13	18333570	62	150006	111	1019
14	14935003	63	139058	112	970
15	12258508	64	129221	113	912
16	10202699	65	119681	114	864
17	8573066	66	110381	115	805
18	7264372	67	102229	116	753
19	6218908	68	94091	117	699
20	5369355	69	86393	118	650
21	4675034	70	80303	119	594
22	4097805	71	73823	120	549
23	3605851	72	67822	121	501
24	3190623	73	61837	122	457
25	2835675	74	55925	123	407
26	2527765	75	50671	124	367
27	2264006	76	46620	125	326
28	2034741	77	42703	126	290
29	1832994	78	38942	127	256
30	1654104	79	35706	128	224
31	1501481	80	32828	129	202
32	1365212	81	30095	130	180
33	1246347	82	27410	131	160
34	1142574	83	24610	132	146
35	1047084	84	22269	133	132
36	959521	85	19805	134	116
37	882049	86	17610	135	102
38	812927	87	15513	136	90
39	749753	88	13430	137	78
40	694949	89	11500	138	66
41	643461	90	9784	139	56
42	596632	91	8311	140	48
43	556771	92	6993	141	42
44	520540	93	6147	142	34
45	487111	94	5418	143	28
46	457686	95	4711	144	22
47	428454	96	4184	145	16
48	400482	97	3699	146	4

It should be noted that 0.48% of the atoms bear a charge. Of these charges, 57.7% are positive and localized (55% are +1); 4.9% are also positive, but delocalized; 34.7% are negative and localized (34.5% are −1); while 2.7% are negative and delocalized. The complete distributions are given for the number of atoms with a given connection degree in Table V, and with a given excentricity in Table VI.

It should also be noted that the values of the excentricities do not depend on bond multiplicity, but would be increased by one unit for most organic compounds if hydrogens were recorded. The excentricity of an atom x varies from R (centroid excentricity) to D (extremal atom excentricity) and is also the maximal number of concentric layers focused around x.

DISTRIBUTIONS IN COMPONENTS

The component is the most useful topological unit defined in a large set of graphs because there is no connection between two atoms which belong to different components. The following statistics for components are derived from a database of 4019514 compounds.

	mean	SD	local maxima	local minima
atoms (without H)	19.384	13.343	1;5;16	2;6;17
hydrogens (mol formula)	17.640	14.408	0;1;3;14;16	2;5;15
carbons	14.312	10.402	10;12;14	11;13
bonds	20.445	14.670	0;4;16	5
cycles (see also ref 2)	2.061	1.899	1;3	2
min connection degree	0.937	0.365	1;4	3
max connection degree	3.042	1.157	3;10	7
radius of components	5.090	3.193	0;5	1
diameter of components	9.634	6.258	0;2;10	1;3
centroids	1.825	1.410	1	
extremal nodes	3.144	1.743	2	

Complete distributions are given for atoms in Table VII, for carbons in Table VIII, for bonds in Table IX, for cycles in Table X, for minimum and maximum connection degrees (bivariate) in Table XI, for radii of components in Table XII, for their diameters in Table XIII, for multiplicities of centroids in Table XIV, and for multiplicities of extremal atoms in Table XV.

The distribution of hydrogen atoms in components shows the same parity phenomenon as it did in compounds, except for components containing two hydrogen atoms, where a local

Table XVIII. Number of Layers with a Given Number of Atoms

atoms	layers	atoms	layers
1	274 232 111	25	588
2	265 064 075	26	1 340
3	175 572 167	27	818
4	96 901 365	28	2 452
5	45 178 432	29	118
6	23 444 257	30	435
7	9 423 141	31	219
8	4 358 166	32	459
9	1 866 522	33	80
10	910 094	34	12
11	397 662	35	38
12	312 082	36	117
13	102 025	37	13
14	101 255	38	4
15	56 960	39	295
16	31 880	40	26
17	16 824	41	210
18	28 884	42	281
19	7 251	45	8
20	7 253	54	222
21	9 610	56	6
22	2 177	60	2
23	670	78	1
24	5 209		

minimum is observed, and components with three hydrogen atoms, where a local maximum is noted. This may be a consequence of the existence of numerous components such as HCl or OH^-, which have only one hydrogen, or those such as NH_3 or H_3PO_4, which have three hydrogens. If components with 0, 1, 2, or 3 hydrogens are removed from the analysis, the relative difference between even and odd distributions is $W_e - W_o = 17.70\%$, which is very close to the 17.57% observed for compounds.

The distribution of carbon in components, shown in Figure 4, shows the same parity phenomenon as that in compounds. In components, the relative difference between even and odd distributions is $W_e - W_o = 7.99\%$ (483 571 components with no carbon were not included). This difference is more than the 6.04% observed for the distribution in compounds. A similar explanation is thought to apply, but the phenomenon may be less pronounced in multicomponent compounds. If one considers an infinite population of components and makes the following assumptions:

(a) one even component has a constant appearance probability p, and one odd component has a constant appearance probability $q = 1 - p$

(b) k-multicomponents are built at random with k independent components

then for each k-multicomponent there is a probability $p(k)$ that it will be even and a probability of $q(k) = 1 - p(k)$ that it will be odd, with $p(1) = p$ and $q(1) = q$ and

$$p(k) - q(k) = (p - q)^k$$

Since $(p - q) < 1$, then $k > 1 \Rightarrow [p(k) - q(k)] < (p - q)$. This means that multicomponent compounds will have a parity difference that is lower than that for monocomponent compounds.

Neglecting the set of compounds with components that are devoid of carbon and using $p - q = 7.99\%$ in conjunction with the data in Table II, these assumptions lead to a bias $W_e - W_o = 6.78\%$, which corresponds to 232 229.8 compounds. The true percentage is 6.04%, corresponding to 204 655 compounds, and so the observed parity is less than the calculated parity. This residual discrepancy is due either to the presence in the Table II data of compounds with carbon-free components or to the inadequacy of the assumptions that have been made. The repartitions of each of the other elements in components have been made and were found to be close to the corresponding distributions in compounds.

The atom distribution, i.e., the number of components with a given number of atoms, excluding hydrogens, has a weak bias $W_e - W_o = 2.8\%$ (403 027 monoatomic compounds being excluded), and this may be explained as the atom distribution concerning the compounds.

The number of components with one atom (0 bonds, 0 cycles, $R = D = 0$, only 1 extremal node) is about 10% of the total number of components. Some 47% of these monoatomic components are either Cl^- or HCl.

The extreme value of the diameter $D = 146$ (Table XIII) is given by two polyethers (RN 1319-63-7 and 42892-32-0) whose structures are shown in Figure 5. The extreme value of the radius $R = 73$ (Table XII) is also due to these two compounds, together with the polyimine (RN = 65000-54-6), which is also shown in Figure 5. The extreme value of the number of centroids (Table XIV) is due to this same polyimine, which has 140 centroids. The extreme value of the number of extremal atoms (Table XV) arises from the macrocyclic compound $(CH_2)_{136}$ (RN = 63217-83-4), which has 136 extremal atoms and the same number of centroids.

DISTRIBUTIONS OF PATHS AND CONCENTRIC LAYERS

Although many processing algorithms that analyze the chemical environment of a focus need to be statistically optimized, there is a paucity of data pertaining to environments. The following statistics refer to the 2 225 906 690 shortest paths from one atom (x) to every other atom (y). There are $n*n$ paths in a component with n atoms and the length of a specific path is the distance d(x,y).

	mean	SD	local maxima	local minima
path lengths	7.489	7.799	3	

The complete distribution is given in Table XVI. The number of paths with zero length is the same as the number of atoms in the whole file. All other path lengths are even values because both distances d(x,y) and d(y,x) are counted.

The following statistics concern the 898 037 816 concentric layers. Each atom x in each component can be viewed as the origin of a set of concentric layers, each y atom in a layer being at a given distance d(x,y) from the atom at the origin.

	mean	SD	local maxima	local minima
distances from origin (depth of layers)	7.286	7.956	1	
atoms in layer	2.479	1.526	1	

The two complete distributions are given in Tables XVII and XVIII. Again, the zero distance corresponds to the number of atoms in the whole file. The number of concentric layers is about half the number of paths because all the distances from an origin to all the atoms that constitute a layer are counted as one layer depth. The number of layers with increasing numbers of atoms does not decrease monotonically after 17 atoms. The extreme value of 78 atoms is due to the disilylmethane (RN = 30859-25-7) shown in Figure 5.

MULTIVARIATE DISTRIBUTIONS

Many bivariate or trivariate statistics are not reported here because the appropriate graphical definition is too large to be conveniently represented. Among the bivariate distributions in components, those for the atoms–bonds, atoms–cycles, bonds–cycles, radius–diameters, and centroids–extremal nodes were computed. The first three of the bivariate distributions are the three projections of the atoms–bonds–cycles distribu-

tion, which lies in a plane defined by

atoms − bonds + cycles =
no. of components of the graph = 1

None of these distributions is readily printed. To do so graphically would require a diagram 70 × 70 cm, and an alphanumeric description would require hundreds of value–value–frequency 3-tuples. The bivariate distribution of the number of layers with a given number of atoms at a given depth has also been computed and presents the same display problems.

All of these bivariate distributions show different local maxima and minima, which define clusters of chemical structures and largely nonexistent chemical entities. It is hoped that multivariate exploration will reveal much more than the univariate or bivariate distributions. For example, the atom–bond–cycle–radius–diameter–centroid–extremal nodes distribution would be expected to show many local extrema, which could be interpreted as families of chemical structures in a statistical classification of the components. The problem is to build an algorithm in a multidimensional space from which the number of groups and their shapes can be computed. This problem is far from a solution.[7]

CONCLUSION

Interpretations have been offered for some of the distributions that have been determined experimentally in this paper, but many of them show unexpected local maxima or minima, or exhibit unusual parity phenomena. It seems clear that the composition of the database cannot be understood simply. Some explanations of the parity phenomena have been advanced, but a full interpretation of the various local extrema remains rather difficult.

The formulation of many algorithms dealing with large data sets can be optimized by means of statistical considerations, for example, substructure searching strategy or organized computerized chemical libraries. Searching and organized substructure data spaces could be achieved advantageously by consideration of the statistical weight of the data and avoiding use of models with poorly defined working spaces. Such working spaces are usually defined pragmatically or by trial and error for specific applications. The predictive space of many quantitative structure–activity relationship endeavors, for example, is usually vague outside the original training set.

The statistics reported here also represent a first step toward a better understanding of the composition of the complete *corpus* of organic chemistry. They show the salient features that are encountered in derivation of an organic chemical classification based upon statistical and topological methods, rather than nomenclatural concepts which stem from the chemist's perception of small sets of data (e.g., the concept of functional groups). Such a classification could lead to new tools for computer-managed structure elucidation and computer-aided property correlation, because the topological approach to molecular structure is more closely bound to chemical and spectroscopic properties than are classical nomenclature systems.

REFERENCES AND NOTES

(1) Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76.
(2) Stobaugh, R. E. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180.
(3) Petitjean M.; Dubois, J. E. *Collect. Czech. Chem. Commun.* **1990**, *55*, 1404.
(4) Dubois, J. E.; Panaye, A.; Attias, R. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74.
(5) Berge, C. Graphes et Hypergraphes. Dunod Université, Paris, 1973. ISBN 2-04-009755-4.
(6) Petitjean, M. Unpublished results.
(7) Everitt, B. S.; Hand, D. J. Finite Mixture Distributions; Monographs on Applied Probability and Statistics. Chapman & Hall: London, 1981. ISBN 0-412-22420-8.