Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-based File. Part III.[1] Statistical Association of Fragment Incidence

By G. W. Adamson, Diane R. Lambourne, and M. F. Lynch,* Postgraduate School of Librarianship and Information Science, The University of Sheffield, Western Bank, Sheffield S10 2TN

The statistical association of the incidence of some pairs of bond- and atom-centred structural fragments has been investigated in a sample of chemical structures from the Chemical Abstracts Service Registry System. Both positive and negative values of the association coefficient have been found, and in general, association is small between atoms but larger between simple pairs, bonded pairs, and augmented atoms. The performance of a structural fragment as a screen in substructure searching is affected by association.

THE cost of searching files of chemical structures is reduced considerably by the use of screens. One of the factors which influence screen performance is the incidence of the structural fragments used as screens in the file of chemical structures being searched. The distribution of some bond-centred [2] and atom-centred [1] structural fragments in a sample file of structures taken from the Chemical Abstracts Service Registry System has been reported and the effect of incidence on screen performance discussed. Many of the terms and symbols used in this paper are explained in these previous publications.

However, most of the possible structures and substructures which could occur as queries contain more than one screen fragment. The prediction of the performance of a set of screens in the case of a query which contains more than one fragment will also depend on the degree of dependence between the incidence of the fragments. If the incidence of the fragments were independent the prediction would be straightforward. For example, if a query were composed of n structural fragments, the ith fragment occurring in a proportion p_i of the structures of the file, and if all n fragments were used as screens, the resultant screen-out would be $100(1 - p_1p_2 \ldots p_n)\%$. If the incidence of the screening fragments were dependent this simple relationship would not be valid and an allowance for association would have to be made.

The results below were obtained in a study to determine the significance of association in the cases of some

[1] Part II, G. W. Adamson, M. F. Lynch, and W. G. Town, *J. Chem. Soc.* (*C*), 1971, 3702.

[2] J. E. Crowe, M. F. Lynch, and W. G. Town, *J. Chem. Soc.* (*C*), 1970, 990.

pairs of the small structural fragments whose incidences have been reported and which have been used as screens in substructure search.

EXPERIMENTAL

Method.—For each calculation a set of m fragments of high incidence was chosen, and records representing these fragments were stored in the computer in the form of a list. One additional word of core store was associated with each item in the list. These additional words were handled as a vector. Each structure in the sample to be studied was read from magnetic tape as a nested, non-redundant connection table and converted into a non-nested, redundant connection table. Fragments were generated from the redundant connection table and compared with the list. When a fragment in the list was found to be present in the structure being analysed the word associated with that fragment was set equal to unity. Thus for each structure an incidence vector I_s (1) was set up. In I_s, $S_{i.s}$ is equal

$$I_s = (S_{1.s}\ S_{2.s} \ldots S_{m.s}) \qquad (1)$$

to unity if the ith fragment, F_i, is present in the sth structure, otherwise $S_{i.s}$ is zero.

The incidence vector I_s was multiplied by its transpose $I^{\dagger}_s$ to give a matrix C_s (2). The diagonal elements c_{ii}

$$C_s = I_s I^{\dagger}_s \qquad (2)$$

are equal to unity if F_i is found in the sth structure and the off-diagonal elements c_{ij} $(i \neq j)$ are equal to unity if F_i and F_j are both present in the structure. Otherwise the elements are zero.

The matrices C_s for each structure were added together as they were formed to give an overall incidence–coincidence matrix E (3) for the sample of m structures being studied.

$$E = \sum_{s=1}^{m} C_s \qquad (3)$$

The diagonal elements e_{ii} are equal to the incidence of the F_i, and the off-diagonal elements e_{ij} $(i \neq j)$ are equal to the number of structures in which F_i and F_j both occur. The elements of E were then divided by the sample size m to give a matrix P in which the diagonal elements p_{ii} are the probabilities of selecting at random from the file a structure containing F_i and the off-diagonal elements p_{ij} are the probabilities of selecting a structure containing both F_i and F_j. If the off-diagonal elements of this matrix were divided by the appropriate diagonal element then conditional probabilities could be obtained: $p_{(j|i)} = p_{ij}/p_{ii}$, where $p_{(j|i)}$ is the probability of finding F_j in the structures which are known to contain F_i. If there is no association between the incidences of F_i and F_j then $p_i \cdot p_{(j|i)} = p_j \cdot p_{(i|j)} = p_i \cdot p_j$. If the incidences of F_i and F_j are associated then $p_i \cdot p_{(j|i)} = p_j \cdot p_{(i|j)} \neq p_i \cdot p_j$. The proportion of structures which contain F_i and F_j is $p_i \cdot p_j$ if their incidences are independent and if F_i and F_j both occur in a query the screen-out would be $100(1 - p_i \cdot p_j)\%$. The actual screen-out observed will be $100(1 - p_i \cdot p_{(j|i)})\%$. The difference between these two values of the screen-out will be due to association. In the case of a query containing n dependent screens the expression for screen-out is $100(1 - p_1 \cdot p_{(2|1)} \cdot p_{(3|1.2)} \cdots p_{(n|1.2.\ldots.n-1)})\%$.

For the comparison of association and its significance between different combinations of fragments the probabilities were converted into association coefficients. The coefficient V which is described by Kendall [3] was used. The expression used to calculate V was (4). V_{ij} is $+1$ if all structures

$$V_{ij} = \frac{p_{ij} - p_i \cdot p_j}{\{(p_i - p_i^2)(p_j - p_j^2)\}^{\frac{1}{2}}} \qquad (4)$$

which contain F_i contain F_j and all structures which contain F_j contain F_i, and V_{ij} is -1 if no structure which contains F_i contains F_j. Otherwise V lies between $+1$ and -1. If there is no association between the incidences of F_i and F_j then $V_{ij} = 0$. V is related to χ^2 by the expression [3] (5). In the results given below the file size

$$\chi^2 = mV^2 \qquad (5)$$

m was 28,831. There is one degree of freedom for a 2×2 association coefficient. Thus for an association coefficient to be significantly different from zero at a 99% confidence level V must lie outside the range $-0{\cdot}02$ to $+0{\cdot}02$.

The association coefficient used here does not distinguish between the direct dependence of the incidence of two fragments and the indirect dependence which may result from dependence between the incidences of these fragments and those of other structural fragments or properties of the structure. Our results are thus useful quantities in the discussion of screen performance but are not very suitable for use in determining the causes of association and apart from one obvious possible cause of association, namely fragment overlap, causality will not be discussed.

The file analysed is a sample taken from the Chemical Abstracts Service Registry System and was described earlier.[2] Compounds containing atoms with a connectivity to non-hydrogen atoms equal to or greater than 5 were omitted. The association coefficients were calculated to 6 decimal places and the values in Tables 1 and 3 were rounded to 2 decimal places.

Programmes were written in PLAN and computation was carried out on the Sheffield University ICL 1907 computer which has a 24-bit word-length and a cycle time of *ca.* 2 μs. The C.P.U. times taken ranged from 1000 s for the calculations for atoms to 2800 s for augmented atoms with bonded pairs. All the computations were carried out in less than $8{\cdot}5 \times 10^3$ words of core store and with one magnetic tape for the input of the CAS nested non-redundant connection tables.

RESULTS AND DISCUSSION

Atoms.—The results for the 15 most frequently occurring atoms are shown in Table 1. The diagonal entries are the numbers of structures in which each atom occurred. The figures in the upper triangle of the Table give the number of structures in which the atoms at the left of the row and at the head of the column both occur. The entries in the lower triangle of the Table are the corresponding association coefficients. The distribution of the 105 association coefficients which were calculated is shown in Table 2.

The co-occurrence of halogen atoms with other halogen atoms and some other heteroatoms in a sample of *ca.* 600,000 structures from the Chemical Abstracts

[3] M. G. Kendall, 'The Advanced Theory of Statistics, Griffin, London, 1943.

Service Registry System has been reported.[4] These results were not given in the form of association coefficients and were converted into this form for comparison with the results shown in Table 1. There were 46 combinations of two atoms which were common to both sets of calculations. In none of these cases was the association coefficient obtained from Leiter and Leighners's results different by more than 0·02 from those given in Table 1.

TABLE 1

The incidences and association coefficients for the 15 most frequently occurring atom types

	C	O	N	S	Cl	F	P	Br	Si	I	B	D	Sn	Se	As
C	28,732	23,787	18,506	5736	4009	2849	1275	1161	399	298	269	127	123	107	95
O	0·03	23,850	15,188	4680	3160	2166	1156	891	244	206	135	70	67	63	60
N	0·06	−0·03	18,524	4366	2668	1548	661	573	83	169	163	43	28	71	37
S	0·01	−0·02	0·12	5752	913	504	364	192	17	31	26	11	30	19	27
Cl	0·00	−0·04	0·02	0·03	4021	522	226	114	92	24	41	6	26	12	13
F	0·01	−0·06	−0·07	−0·02	0·04	2852	98	146	68	52	94	12	9	5	9
P	−0·05	0·04	−0·06	0·04	0·02	−0·02	1296	36	12	6	10	0	3	4	3
Br	−0·01	−0·03	−0·06	−0·02	−0·02	0·02	−0·01	1167	8	6	12	11	6	2	2
Si	−0·03	−0·07	−0·11	−0·05	0·03	0·03	−0·01	−0·01	407	3	6	2	2	0	0
I	−0·03	−0·04	−0·02	−0·03	−0·02	0·02	−0·01	−0·01	0·00	304	3	2	2	1	2
B	−0·02	−0·09	−0·01	−0·03	0·00	0·08	0·00	0·00	0·01	0·00	274	1	2	1	0
D	0·00	−0·05	−0·04	−0·02	−0·02	0·00	−0·01	0·02	0·00	0·00	0·00	127	0	0	0
Sn	−0·03	−0·05	−0·06	0·01	0·01	−0·01	−0·01	0·00	0·00	0·00	0·00	0·00	127	0	0
Se	−0·05	−0·04	0·00	0·00	−0·01	−0·01	0·00	−0·01	−0·01	0·00	0·00	0·00	0·00	113	0
As	−0·01	−0·03	−0·03	0·01	0·00	0·00	0·00	−0·01	−0·01	0·01	−0·01	0·00	0·00	0·00	96

The total number of atom types in the file is 62. However, the incidence of the atoms falls so rapidly with increasing rank that many pairs of different atoms will never be found to occur in the same structure in this file. The statistical association found between atoms in this file in small. Of the 105 association coefficients calculated 92 lie between −0·05 and +0·05 and 103 between −0·10 and +0·10. The two association coefficients which lie outside this range are V(N,S) = 0·12 and V(N,Si) = −0·11. Nitrogen and sulphur occur in the same compound 4366 times and nitrogen and silicon 83 times. If the incidences of these atoms had been independent then these figures would have been 3695 and 261 respectively. Association would thus affect the performance of these pairs of atoms as screens. The negative association between silicon and nitrogen will increase screen-out and the positive association between sulphur and nitrogen will result in a decrease in screen-out.

TABLE 2

Summary of sign and magnitude of associations within and between fragment classes

	Atoms	Simple pairs	Bonded pairs	Augmented atoms	Bonded pairs, augmented atoms
$V < 0·25$	0	0	0	0	0
$-0·25 < V < -0·15$	0	0	6	5	6
$-0·15 < V < -0·05$	10	41	55	36	17
$-0·05 < V < 0·05$	92	106	59	75	32
$(-0·02 < V < 0·02)$	(65)	(35)	(28)	(34)	(11)
$0·05 < V < 0·15$	3	25	37	43	13
$0·15 < V < 0·25$	0	7	16	16	10
$0·25 < V < 0·35$	0	6	4	6	5
$0·35 < V < 0·45$	0	2	7	5	3
$0·45 < V < 0·55$	0	2	4	3	3
$0·55 < V < 0·65$	0	1	0	1	1
$0·65 < V < 0·75$	0	0	0	0	5
$0·75 < V < 0·85$	0	0	1	0	2
$0·85 < V < 0·95$	0	0	1	0	3
$V > 0·95$	0	0	0	0	0
Total number of association coefficients calculated	105	190	190	190	100

Simple Pairs.—The data for the most frequently occurring 20 simple pairs are in Table 3. The diagonal elements give the incidence of the fragments. The figures in the upper triangle of the Table give the number of compounds in which the simple pairs at the left of the row and head of the column of the entry concerned both occur. The entries in the lower triangle of Table 3 are the corresponding association coefficients. The distribution of the association coefficients is shown in Table 2.

Compared with the association coefficients found for atoms, those found for this set of simple pairs cover a wider range although most of them still lie between −0·05 and +0·05. The distribution of the association coefficients calculated is unsymmetrical about $V = 0$, with larger values of positive association found than of negative association.

In Table 3 the simple pairs have been arranged in three classes according to whether they occur in chains, in rings with localised bonds, or in rings with aromatic bonds. At the time the data base was compiled the definition of aromatic used by Chemical Abstracts Service was alternating single and double bonds in a ring. In general the largest positive associations are found between simple pairs which are from the same classes. This is pronounced in the case of simple pairs which contain localised ring bonds where only one of the 15 coefficients calculated is negative and 7 of the coefficients are greater than +0·20. The value of V(C*C,C*N) is also large and positive and the largest positive value of V found is V(C–S,S=O) = 0·63.

[4] D. P. Leiter and L. H. Leighner, ' A Statistical Analysis of the Structure Registry at Chemical Abstracts Service,' presented at the ACS Meeting, Chicago, 1967.

Association between simple pairs belonging to different classes is generally smaller and more often negative than positive.

These observations can generally be explained qualitatively by relating positive association to the overlap between simple pairs. As simple pairs which were generated were centred on every bond in the structures, atoms other than terminal atoms may form part of more than one simple pair. For example, the group C–C(–O)=O found in carboxylic acids, salts, and esters would give three different simple pairs C–C, C–O, and C=O which share one carbon atom and the association coefficients between these simple pairs are V(C–C,C–O) = 0·27, V(C–C,C=O) = 0·27, and V(C–C,C=O) = 0·28. In the cases of simple pairs containing ring bonds more overlap is implied than is explicit in the representation of the fragment because an atom in a ring must form at least two ring bonds. For example in the simple pairs C·C and C·N each atom must be bonded by at least one more ring bond than is shown in the fragment, leading to the possibility of further overlap between fragments.

The largest positive and negative associations found are V(S=O,C–S) = +0·63 and V(C·C,C–S) = −0·15. If their incidences were independent then S=O and C–S would occur in 258 structures and C·C and C–S in 1714, whereas the combinations of fragments actually occur in 1810 and 1034 structures respectively. The positive association between S=O and C–S is explicable in terms of overlap of the fragments which could have a common sulphur atom. However, the negative association between C–S and C·C is not explicable in these terms as these two fragments could have a common carbon atom.

The association coefficients of V(C*C,N=O) = 0·15 and V(C*C,S=O) = 0·12 are also not explicable in terms of direct overlap. However, the simple pairs N=O and S=O occur, for instance, in nitro- and sulpho-groups which are frequently found as substituents in aromatic compounds. The simple pairs C–S and C–N which could be between S=O and N=O and the aromatic ring have positive associations with C*C and with C–S and C–N.

The association coefficients which were calculated for other fragment types are not all given but the summaries of the values found are shown in Table 2 and individual values judged to be of interest are discussed in the text.

Bonded Pairs.—The bonded pair is a larger bond-centred fragment than the simple pair, and spans up to three bonds and two atoms in a structure, whereas the simple pair spans one bond and two atoms. The distribution of the association coefficients for the 20 most frequently occurring bonded pairs is shown in the third column of figures in Table 2. Compared with the corresponding figures for simple pairs the association coefficients are generally larger, and fewer than one third of them lie in the range −0·05 to +0·05. The range of values found is also greater, being from V(*Ċ*Ċ*,C*Ċ*) = +0·91 to V(·C·Ċ·,*C*Ċ*) = −0·19.

If the bonded pairs are split into three classes according to the central bond type as described for simple pairs, then similar trends are apparent. However some bonded pairs have pendant bond types which do not belong to the same class as the central bond, leading to modifications of the basic classification of fragments used for the simple pairs.

Augmented Atoms.—Augmented atoms are atom-centred fragments which span up to three atoms and two bonds in a structure. Some features of the distribution of augmented atoms have been reported.[1] The association coefficients between the 20 augmented atoms which occur most frequently in the file were calculated, and a summary of the results is shown in Table 3.

The association coefficients found range from V[C*C*C,C*C(–C)*C] = +0·60 to V(C*C*C,C·C·C) = −0·20. If these fragments occurred independently C*C*C and C*C(–C)*C would have occurred in 6711 structures and C*C*C and C·C·C in 4847, whereas these combinations actually occurred in 10,891 and 3653 structures respectively. The fragments, when divided into groups according to bond types present, show similar trends in association to those observed in simple pairs and bonded

TABLE 3

Incidences and association coefficients for the 20 most frequently occurring simple pairs

	C–C	C–O	C=O	C–N	C–Cl	C–S	C=C	C–F	N=O	O=S	C=N	O–P	C·C	C·N	C:C	C·O	C:N	C·S	C*C	C*N
C–C	24,423	15,087	14,386	12,950	3060	2845	3137	2384	1908	1763	1863	830	11,999	6336	5235	3752	2379	1403	15,403	2124
C–O	0·27	16,168	10,602	7885	1788	1710	1891	1375	1231	1105	974	854	7903	3622	3461	2658	1164	610	10,131	1286
C=O	0·27	0·28	15,344	8377	1851	1561	1863	1366	954	914	869	342	8777	4345	3976	2673	1307	806	9499	915
C–N	0·11	−0·04	0·08	14,615	2034	2118	1329	1200	2208	1480	1733	405	6795	5596	2666	1803	1985	1029	11,276	1658
C–Cl	0·01	−0·05	−0·01	−0·04	3592	536	376	450	289	344	268	140	1324	876	574	320	361	218	2754	413
C–S	−0·01	−0·04	−0·05	0·08	0·04	3403	253	297	282	1810	289	180	1034	734	403	277	422	248	2603	384
C=C	0·12	0·02	0·03	−0·07	−0·01	−0·04	3236	232	258	145	157	90	1477	652	675	446	236	252	1778	244
C–F	0·04	−0·03	−0·01	−0·04	0·04	−0·01	−0·03	2665	214	203	143	47	934	414	499	187	154	136	1514	147
N=O	−0·02	−0·02	−0·07	0·26	0·00	0·00	0·00	0·00	2327	158	401	42	907	598	443	305	302	116	2026	295
O=S	−0·04	−0·04	−0·07	0·09	0·03	0·63	−0·04	0·00	−0·01	2210	193	40	781	506	267	194	221	243	1830	186
C=N	0·04	−0·08	−0·06	0·19	0·00	0·02	−0·03	−0·02	0·12	0·02	2061	29	862	397	417	227	146	145	1529	123
C–P	0·02	0·13	−0·06	−0·03	0·01	0·04	−0·01	−0·03	−0·02	−0·02	−0·03	932	290	198	134	218	137	13	420	97
C·C	−0·08	−0·04	0·14	−0·09	−0·11	−0·15	−0·04	−0·10	−0·07	−0·09	−0·05	−0·07	14,651	6590	5982	4339	1951	1244	8438	939
C·N	−0·06	−0·12	0·03	0·26	−0·02	−0·05	−0·06	−0·08	−0·01	−0·03	−0·05	−0·02	0·41	7790	2823	1552	2623	1139	5855	811
C:C	−0·03	−0·02	0·10	−0·09	−0·05	−0·09	−0·01	−0·03	−0·02	−0·07	−0·01	−0·03	0·46	0·21	6342	1994	1331	764	3277	301
C·O	−0·05	0·01	0·04	−0·11	−0·08	−0·08	−0·02	−0·08	−0·03	−0·06	−0·04	0·04	0·37	0·06	0·21	4668	739	198	2649	335
C:N	−0·02	−0·11	−0·05	0·12	0·00	0·03	−0·03	−0·04	0·03	0·00	−0·03	0·03	0·11	0·48	0·19	0·09	2886	579	2187	386
C·S	−0·04	−0·11	−0·04	0·04	0·00	0·02	0·02	−0·01	−0·02	0·06	0·01	−0·04	0·10	0·21	0·13	−0·04	0·19	1783	1299	117
C*C	−0·01	−0·02	−0·03	0·29	0·10	0·10	−0·06	−0·04	0·15	0·12	0·06	−0·07	−0·12	0·15	−0·13	−0·06	0·09	0·05	18,925	2598
C*N	−0·05	−0·05	−0·12	0·07	0·03	0·03	−0·02	−0·04	0·03	−0·01	−0·03	0·01	−0·10	0·02	−0·08	−0·03	0·05	−0·02	0·22	2679

pairs. For example, as augmented atoms are generated from every atom in a structure an ester group –C–C(=O)–O–C– would give rise to the following augmented atoms: C–C(–O)=O, O=C, and C–O–C. The association coefficients found between these fragments are V[C–C(–O)=O, O=C] = +0·53, V[C–C(–O)=O, C–O–C] = +0·45, and V(C–O–C, O=C) = +0·27. This association can again be explained in terms of overlap between fragments although these fragments could be generated from groups other than an ester group.

Augmented Atoms and Bonded Pairs.—Augmented atoms and bonded pairs span different regions of a structure. Augmented atoms may be generated from adjacent atoms and bonded pairs from adjacent bonds. Association between different bonded pairs or different augmented atoms is large in some cases and is related in part at least to overlap between fragments. Thus, high association would be expected between bonded pairs and augmented atoms which may be centred on a bond and an adjacent atom, and could have a large region of overlap.

Association coefficients were calculated between 10 of the most frequently occurring bonded pairs and 10 of the most frequently occurring augmented atoms. The results are summarised in the last column of Table 2. Out of 100 coefficients calculated 10 are greater than 0·65. The largest values found are V(C*C*C,*C*C*) = 0·91, V(C–C–O,–C–O–) = 0·87, and V(C*C*C*,*C*$\overset{|}{\text{C}}$*) = 0·85. The largest negative value found was −0·23 for V(C·C·C,*C*$\overset{|}{\text{C}}$*).

The high positive association between *C*C* and C*C*C leads to the following results in the case of a query which contains the structural fragment *C*C*C. If this were run against the file of 28,831 structures used in this analysis and the effects of other fragments present in the query were ignored then, by use of the bonded pair *C*C* alone, 12,374 compounds would be screened out. If the augmented atom C*C*C were used alone then 11,182 structures would be screened out. If *C*C* and C*C*C were used together 12,393 structures would be screened out. The additional screen-out gain by using both the augmented atom and the bonded pair compared with that using only the bonded pair is 19 structures in 28,831, *i.e.*, less than one structure per thousand. This is the most extreme case of association which we have found and in every case investigated the gain by using augmented atoms and bonded pairs would be greater. However, this is an example in which, because of association, very little is gained by using a combination of fragments in a query.

Conclusions.—The incidences of the structural fragments considered in this work are not independent. Dependence as illustrated by the values of the association coefficients calculated is small in the case of atoms and larger for simple pairs, bonded pairs, and augmented atoms, and larger for bonded pairs than for simple pairs. The largest association was found in the case of some combinations of bonded pairs and augmented atoms. Some association values found are large enough to have a considerable effect on the performance of a screening system in the cases of a system and a query which use that combination of fragments. Both positive and negative associations are found but positive association seems to become more important as larger, more precisely described, structural fragments are used. When a query is made up of two positively associated fragments then association will reduce screen-out and the opposite is true in the case of negative association.

Although the investigation was designed to show the presence and size of association and its effects on screen performance and not the causes of association, many of the observed association coefficients can be explained qualitatively in terms of the possibilities of overlap between fragments in structures. In this work and in some screening systems fragments are generated from all possible centres of a particular type, bonds, atoms, or rings in a structure. Thus, once fragments have reached a certain size, those generated from adjacent centres will overlap and the region of overlap will increase as the size of the fragments is increased further. If fragments which do not overlap are used as screens then these effects can be reduced but the number of possible fragments in a structure which may be used as screens will be reduced and difficulties will be encountered in the case of queries which do not exactly match any combination of the non-overlapping screening fragments.

If a hierarchical set of overlapping fragments is used then queries can be represented for search accurately and simply, with small generally defined structural fragments acting as broad terms and larger precisely defined structural fragments as narrow terms. Several fragments of different size would then be developed from each centre. For example, a simple pair, an augmented pair, and a bonded pair could be developed by use of the same bond in a structure as a centre. The parts of a query which were well defined could then be coded for search by use of large fragments and the less precisely defined parts of a query by use of the small fragments. In this case the association between fragments of different size (*e.g.*, between simple pairs and bonded pairs) would also affect performance. The use of a hierarchical set of fragments introduces additional redundancy into the fragment representation of a structure, but this redundancy is of use in easing query coding.

The associations observed would have to be allowed for in an exact model of the behaviour of a set of screens or in a screening system which was designed to give optimum performance. However, screening systems capable of giving high average screen-out seem to have been designed without much consideration being given to association and it seems that for screening systems which use fragments at least as large as bonded pairs or augmented atoms, association may be largely ignored

at the practical level, although it is important from the theoretical point of view. If larger screening fragments are used then it seems likely to become more important, and because of it the use of some large structural fragments may be less rewarding than otherwise appears.

The association between structural fragments would also have to be taken into account in accurate calculations of association between the presence or absence of structural fragments and other molecular properties.

We thank the Office for Scientific and Technical Information for financial support and Chemical Abstracts Service for providing a file of chemical structure in machine readable form.

[2/854 *Received*, 17*th April*, 1972]