SLASH: A program for analysing the functional groups in molecules

D. A. Cosgrove* and P. Willett†

**Zeneca Pharmaceuticals, Macclesfield, Cheshire, UK*
†Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield, UK

The program SLASH, which is designed to enable large numbers of compounds to be analysed in terms of the functional groups they contain, is described. The usefulness of the groups for analysing the activities of compounds tested in high-throughput biological screens is investigated. The functional group fragments are also studied as a means of determining similarity within groups of compounds. © 1998 by Elsevier Science Inc.

Keywords: molecular similarity, high-throughput screening, functional groups

INTRODUCTION

The advent of high-throughput screening (HTS) has created a situation in which large numbers of compounds are tested, with a correspondingly high number of active compounds identified. This has placed an increasingly onerous burden on chemists receiving the results at the end of a testing run, who are required to assess rapidly the active compounds in order to nominate potential lead compounds. This article describes work carried out to analyse large collections of compounds in terms of the functional groups they contain, with a view to identifying those functional groups linked to biological activity.

A number of fragment descriptors have been used in an effort to explain the activity of compounds in biological tests. Some of the commonest descriptors include database search keys,[1–3] augmented atoms and their derivatives,[3–5] topological torsions,[6] molecular fingerprints,[‡] fragments automatically derived from Wiswesser codes (i.e., Wiswesser line notation, WLN),[8–11] and small databases of fragments.[12–14] This has generally been carried out with a view to predicting properties, such as biological activity or toxicity, of untested compounds. The work described herein was aimed at the somewhat different task of explaining the activity of tested compounds in terms that would be acceptable to a chemist trying to identify a lead compound or series. None of the preceding descriptors are particularly useful for this, except, perhaps, for the fragments derived from WLN. The database search keys, for instance, are designed to optimise the searching of databases. There is no requirement for the fragments represented by the keys to make any sort of chemical sense, and, frequently, rare but quite different substructural species are allocated to the same key, which further complicates the analysis. Augmented atoms are rather too small to be useful in isolation, consisting as they do of only three or four atoms and associated bonds. There are also problems with the superimposed coding methods used for creating the molecular fingerprints, such as those in the Daylight software,[7] since here it is impossible to go from the fingerprint to the fragments represented by that fingerprint. Thus, whilst such fingerprints are extremely useful for grouping together similar compounds, there is no way to identify automatically what it is about those compounds that is deemed to be similar, although the Stigmata program goes some way towards this.[15] The WLN fragments are generally recognisable functional groups, but this method of molecular notation is falling into disuse, and there are technical problems associated with the automatic fragmentation process.

The program SLASH was written to address these issues. It uses MACCS SD files as input, so that it is relatively convenient to prepare files for analysis. The fragments produced are generated algorithmically from the molecules in the input set. Approaches that use databases of previously defined fragments suffer from the restriction that they are only as good as the fragment database. There is always the possibility that particular groups of importance may be missing from the fragment database, causing the user to miss useful correlations, unless the fragment database is tailored to the set in question. The algorithmic approach, whereby fragments are generated from the molecules input, does not suffer from this restriction. It may be the case that the rules used to generate the fragments miss some fragments that might be of interest, but with a reasonable set of rules this should be minimised.

Once the fragments have been generated, there remains the problem of analysing them, in an attempt to identify those functional groups that are predominantly found only in active or inactive compounds. In the ideal case, it would be possible

[‡] For example, the Daylight (Daylight Chemical Information Systems, Inc., 419 East Palace Avenue, Santa Fe, NM 87591) and Unity (Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2913) fingerprints

Address reprint requests to: D.A. Cosgrove, Zeneca Pharmaceuticals, Mereside, Alderly Park, Mecclesfield, Cheshire SK10 4TG, UK.

Accepted 21 May 1998.

Journal of Molecular Graphics and Modelling 16, 19–32, 1998
© 1998 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/98/$19.00
PII S1093-3263(98)00014-X

to group the molecules into sets according to which functional groups they contain, and those sets would contain mostly active or inactive compounds. This is rarely possible in practice, so SLASH aims to give some measure of the probability that a particular fragment is implicated in the activity. To this end, some means is required to give each fragment a score that reflects its presence in active or inactive compounds, a process frequently termed *substructural analysis*.

A number of substructural analysis scoring functions have been proposed and compared previously,[16] with the best of these appearing to be the R2 relevance weight of Robertson and Sparck-Jones.[17] This assessment is based on the investigation of how well the activities of untested compounds are predicted by summing the scores (or weights) for the fragments in a molecule when the weights have been previously calculated for another or the same set of molecules with known activities. In previous studies, the predictions have not been overly impressive, although they have been shown to be statistically significant. One factor that has been mentioned[1] as a potential cause of the less-than-perfect performance has been the fragments themselves, which have generally been database search keys or some of the smaller fragment types detailed above. One aim of the SLASH study was to see if the use of more sensible functional group fragments in a substructural analysis could improve the performance of any of the fragment scoring functions.

THE SLASH FRAGMENTATION RULES

SLASH takes a set of molecules, and derives from them a set of fragments according to predefined rules. When establishing the rules, the following considerations were taken into account:

- The fragments are created from the molecule set in question, allowing for an open-ended set of descriptors.
- The fragments should, if possible, comprise "sensible" chemical species. Some of the automated fragmentation procedures that have been reported in the literature[18–19] can produce fragments that are not sensible in isolation, such as portions of an aromatic ring. The database keys used in many of the studies of fragment weighting schemes referred to above are particularly prone to this criticism, since they are usually selected solely on their statistical properties in test databases and can contain some rather odd groupings of atoms.
- When forming the fragments in a particular molecule, subfragments of larger entities (such as methylene groups that are parts of longer aliphatic chains) are not reported explicitly. This helps to control the number of fragments generated and reduces the amount of redundant information.
- An atom in a particular molecule may appear in more than one fragment, but only if those fragments are in different classes. (The next section describes the different classes of fragment.)
- There is no upper limit on fragment size. This can lead to some quite large fragments, but it precludes the problems that can be seen when such a limit is imposed. Topological torsion fragments,[6] for instance, are four-atom descriptors comprising the four connected atoms that might be used to describe a torsion angle. When faced with the structure O=CNC=O (SMILES notation), such descriptors would completely miss the presence of the two carbonyl groups. The absence of any maximal fragment size also means that under no circumstances are aromatic rings cleaved to leave a partial ring structure.

Clearly, no set of rules for fragmenting molecules can capture the full complexity of chemical possibilities, or even all the possibilities likely to be of interest to a medicinal chemist. The rules used in SLASH were the results of discussions with medicinal chemists at Zeneca Pharmaceuticals (Macclesfield, UK) and therefore reflect their particular biases. There are bound to be arbitrary decisions that produce slightly less than ideal results under some circumstances, but it is hoped that these have been kept to a minimum.

The different fragment classes

Three classes of fragments have been defined: rings, groups, and chains. They are defined below.

Rings Any cyclic system. The definition is extended to include one-atom unsaturated systems that are exocylic to an aromatic ring. Thus the =O group of pyridone is included as part of the ring fragment.

Groups Collections of noncarbon, nonhydrogen atoms that a chemist might normally consider as "functional groups" or "substituents." The definition is extended to cross an unsaturated carbon that is part of a larger group, to ensure that groups such as amides and esters are included as one fragment.

Chains Any chain of acyclic carbon atoms. Saturated and unsaturated systems are treated as equivalent during the fragmentation process, all other things being equal.

The classes of fragments just described could well produce a set of fragments that were too detailed to be useful, with consequent statistical problems occasioned by the resulting low frequencies of occurrence. If each fragment occurs in only one or two compounds, it would be difficult and probably foolhardy to try and draw any statistical conclusions from their distributions. To tackle this problem, a hierarchy of fragments was generated, with the fragment being at one end extremely general, and at the other end precisely differentiated.

The hierarchical fragmentation approach

Each class of fragment is divided into a hierarchy that allows the comparison of fragments at various levels of complexity. Thus two fragments that are different at one level might be equivalent at a lower level. With the ring fragments, for instance, it may be the case that there are insufficient examples of furans and pyrroles to allow firm conclusions to be drawn, so comparisons could instead be made by grouping them together as aromatic five-membered heterocycles, or even simply as rings. It is this sort of possibility that the hierarchy is intended to address. Each fragment class has its own hierarchy, reflecting the sorts of comparisons that are likely to be required.

The ring hierarchy The ring hierarchy is the most detailed hierarchy, with the greatest number of levels. Reference is made to Figure 1, which shows a number of different ring fragments that will be used as examples. Rings that are equiv-

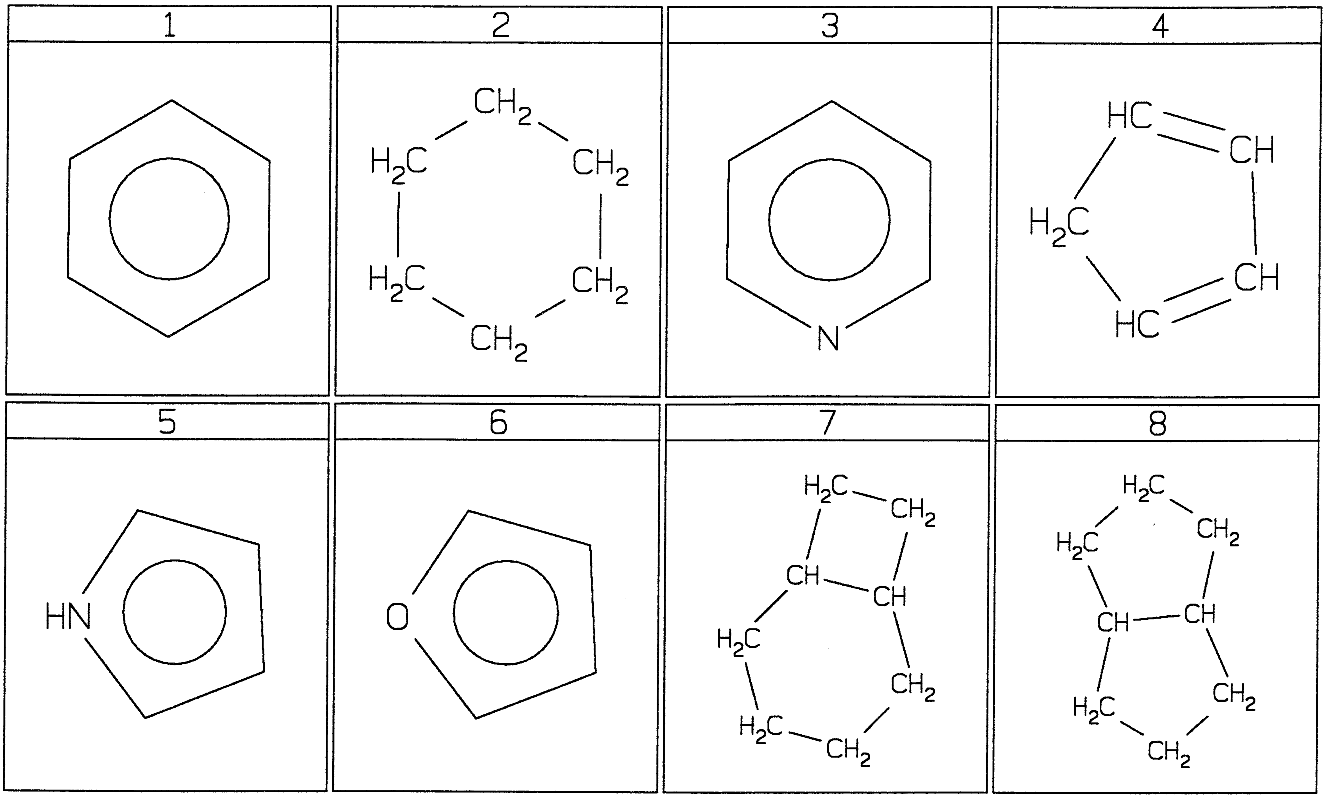

Figure 1. Different ring fragments.

alent at the level being considered will be grouped together in braces at the end of the level description.

Level 0 Level 0 describes just a ring. It thus flags a molecule as having one or more cyclic fragments. {1 2 3 4 5 6 7 8}

Level 1 Level 1 distinguishes rings by the number of atoms they have. {1 2 3}, {4 5 6}, {7 8}

Level 2 Level 2 distinguishes rings according to connectivity patterns, ignoring atom types, but considering bond types. {1 3}, {2}, {4 5 6}, {7}, {8}

Level 3 Level 3 distinguishes rings according to connectivity patterns and atom types, except that all noncarbon, nonhydrogen atoms are considered equivalent. It thus differentiates between homocyclic and heterocyclic rings. {1}, {2}, {3}, {4}, {5 6}, {7}, {8}

Level 4 Level 4 distinguishes between rings according to connectivity patterns, taking into account both atom and bond types. {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}

The group hierarchy The group hierarchy has four levels, defined below. The example groups are shown in Figure 2. Connection points are indicated by the xenon atoms.

Level 0 Level 0 describes just a group. It flags the molecule as having one or more noncarbon heavy atoms. {1 2 3 4}

Level 1 Level 1 differentiates between groups according to the number of atoms they contain. {1 2}, {3 4}

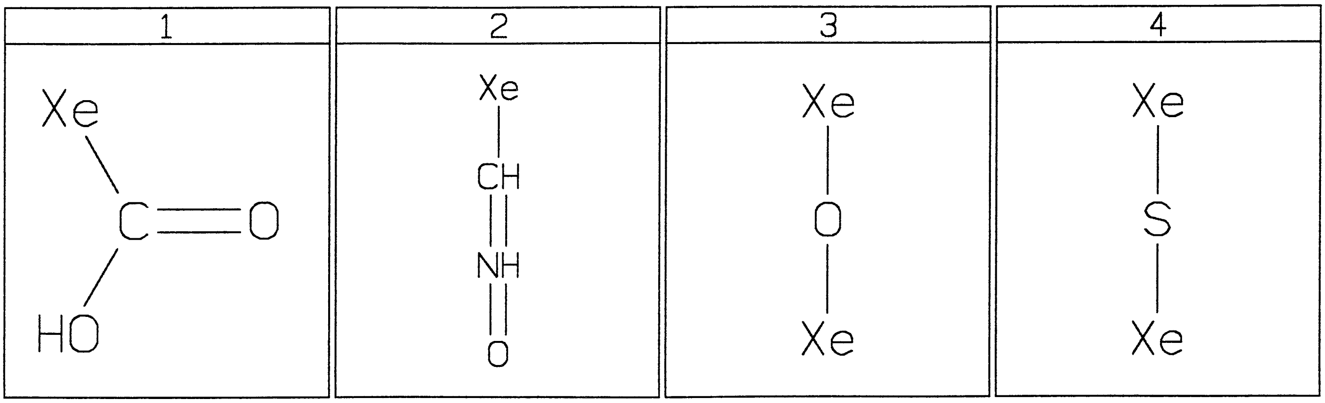

Figure 2. Different group fragments.

Level 2 Level 2 differentiates between groups with the same number of atoms, but different connectivity patterns, ignoring atom types. {1}, {2}, {3 4}

Level 3 Level 3 differentiates between groups according to both connectivity patterns and atom types. {1}, {2}, {3}, {4}

The chain hierarchy The chain hierarchy is the simplest hierarchy, as there are fewer possibilities. Examples are given in Figure 3, which also has xenon atoms marking the connection points.

Level 0 Level 0 describes just a chain of carbon atoms. {1 2 3 4}

Level 1 Level 1 distinguishes between chains with different numbers of atoms. {1}, {2}, {3 4}

Level 2 Level 2 differentiates between chains with the same number of atoms, but different connectivity patterns. {1}, {2}, {3}, {4}

An example of the type of fragmentation pattern produced by SLASH is given in Figure 4, which shows analysis of the molecule MCMC00000019, taken from the MACCS Comprehensive Medicinal Chemistry database. The full structure is drawn in the top left-hand corner of Figure 4, with the fragments shown ringed in the remaining pictures. Only one ring is found: fragment 1. There are three groups (fragments 2–4) and four chains (fragments 6–9). Fragments 8 and 9 are identical at all levels, the others being unique at their highest level of comparison.

Implementation notes

When performing an analysis of a set of molecules, each molecule is fragmented according to the above-described rules. Each fragment is then checked to see if it has been found already in a previous molecule. If it has, the list of molecules associated with that fragment is updated. If not, the new fragment is added to the set of all fragments.

The program is implemented using the C++ programming language, with the graphical interface written using Motif. This allows the user to view all of the molecules and the fragments. In addition, it is possible to select a fragment and view in a separate window the molecules that contain that fragment. Fragment scores, as discussed in the next section, may also be calculated and displayed, so that the fragments can be viewed in ascending or descending order of score. This allows the user to see which fragments are perceived to be the most beneficial for activity, and which are the least beneficial. It is possible to exclude fragments from the weighting scheme if they occur in only a few molecules, thus enabling the user to concentrate only on those fragments that are widespread in the data set.

FRAGMENT WEIGHTING SCHEMES

All the fragment weighting or scoring schemes that have been reported in the literature[16] involve some or all of the four following pieces of information about a fragment:

$NACT_i$ The number of active compounds with fragment *i*.
$NTOT_i$ The total number of compounds with fragment *i*
NACT The number of active compounds in the data set
NTOT The total number of compounds in the data set

Four such scores were investigated for analysing the distribution of SLASH fragments in various sets of molecules, these being the ones that Ormerod et al. found gave the best results.[16,20]

Cramer Structure Activity Fraction (SAF) weight

The SAF weight of Cramer et al.[1] is the simplest weight, and is given by the formula

$$\mathrm{SAF}_i = \mathrm{NACT}_i/\mathrm{NTOT}_i$$

It is generally the poorest of the weights, most probably because it takes no direct account of the number of inactive compounds that contain the fragment.

Cramer Structure Activity Score (SAS) weight

The SAS weight of Cramer et al.[1] has the formula

$$\mathrm{SAS}_i = \mathrm{NACT}_i - (\mathrm{NTOT}_i \times \mathrm{NACT/NTOT})$$

It behaves in a more consistent manner than the SAF weight, but less well than the two described below. It is a number that represents the difference between the number of active compounds containing the fragment in question, and the number that could be expected if the fragment had no influence on the activity of the molecule.

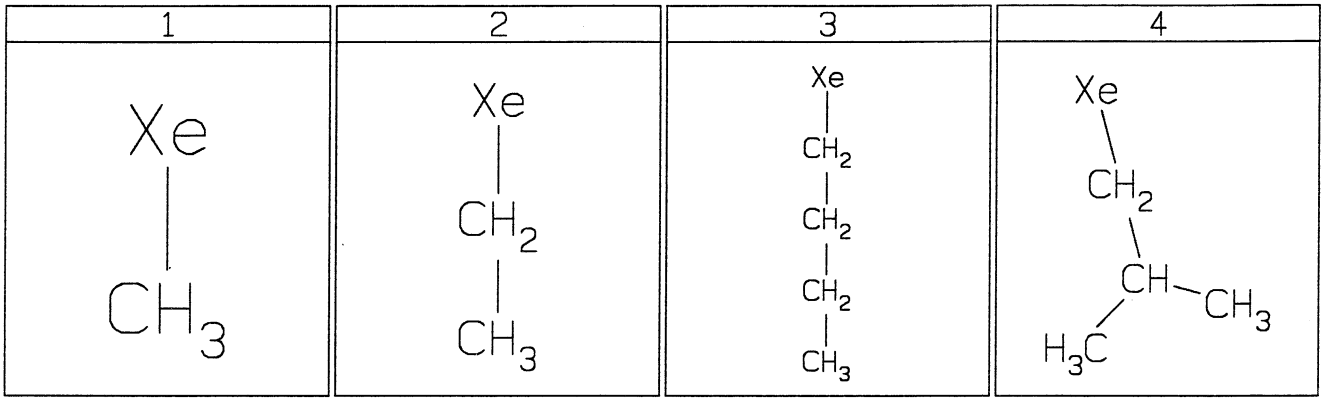

Figure 3. Different chain fragments.

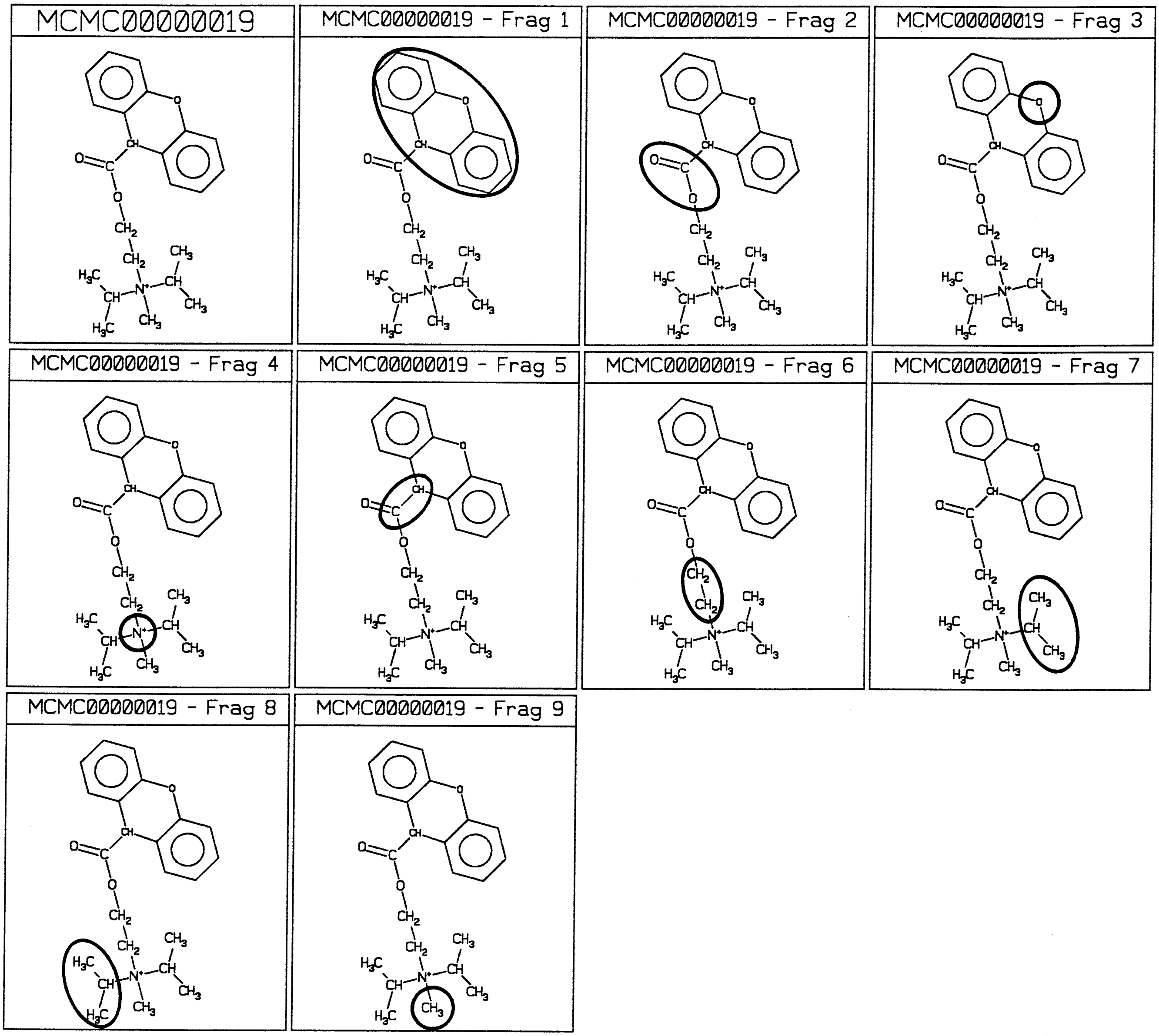

Figure 4. Fragmentation of MCMC00000019.

Hodes Number of Standard Deviations (NSD) weight

The Hodes NSD weight[3] cannot be described by a simple formula. The number of active compounds containing fragment *i* is compared with the number of fragments that would be expected if there were a binomial distribution of fragments among molecules. The number of standard deviations by which the one differs from the other is used as the fragment score. This measure performs comparably to the Robertson and Sparck–Jones weight in some cases, but is marginally less consistent overall.

The Robertson and Sparck–Jones R2 weight

The Robertson and Sparck–Jones R2 weight[17] has the formula

$$R2 = \log[(NACT_i/NACT)/(NINACT_i/NINACT)]$$

where $NINACT_i$ is the number of inactive compounds with fragment *i*, and NINACT is the number of inactive compounds in the data set. While less intuitively obvious than either of the Cramer weights, the R2 score is the result of a detailed probabilistic analysis of the problem. Ormerod et al. found that it gave the best overall performance of all the fragment weights described and investigated in the literature, and this was borne out by the results with SLASH.

THEORETICAL COMPARISON OF THE FOUR SCORES

Cosgrove has carried out a theoretical analysis of the performance of the four scoring functions described.[21] An artificial data set was created comprising 200 compounds, half of which were deemed active, the other half inactive. For each of the four scoring functions, values were calculated for the cases in which a fragment was found in zero to nine active and zero to nine inactive compounds, giving, for each score, a matrix of 100 values. Thus, given a fragment occurring in, for example, 3 of the 100 active molecules and 7 of the 100 inactive molecules, the four weights would have the following values: 0.3 (SAF), −2.0 (SAS), −0.92 (NSD), and −0.37 (R2).

Four criteria were used to perform a qualitative evaluation of the scores:

1. The scores should be able to distinguish between fragments that appear to promote activity, and those that appear to be detrimental.
2. The scores should take account of the total number of active or inactive compounds in the set, as well as the total number of active or inactive compounds with a particular fragment.
3. The scores should incorporate information about the difference in the number of active and inactive compounds containing the fragment. Thus, a fragment that appears in 15 active and 1 inactive compound should receive a higher score than a fragment that appears in only 1 active compound, and no inactive ones.
4. For a hypothetical data set containing equal numbers of active and inactive compounds, the absolute values of the scores for activating and inactivating fragments should be the same. For instance, if a fragment occurred in nine of the active compounds, and in none of the inactive ones, the R2 value is 7.95. If the fragment occurs in nine of the inactive compounds and in none of the active ones, its value is −7.95. On the other hand, the SAF weight gives 1.0 in the former case (nine of nine) and 0.0 in the latter (zero of nine) because $NINACT_i$ does not appear in the formula. NSD and SAS also pass this test because, whilst $NINACT_i$ does not appear explicitly in the calculation, it is implied by the fact that $NTOT_i = NACT_i + NINACT_i$.

On the basis of these criteria, Cosgrove found the SAF score to be the least useful; R2 was next best, with little to choose between SAS and NSD as the best. The SAF score fails on criterion 4, and the R2 weight suffers because of the way in which it treats fragments that are found solely in active or inactive compounds. In these cases, to avoid a division-by-zero problem, a value of 10^{-7} was used instead of zero. This results in spuriously large values in these cases. The fact that R2 appears to work best in practice despite this limitation could be because in real data sets it is relatively unusual for a fragment to occur in only the active or inactive molecules. The problem is more likely to arise in the current work, however, since the automated fragmentation process often produces fragments that are seen only in one compound in a set, a situation that is not likely when the fragments being used are database search keys, which, by their nature, turn up in a reasonable percentage of molecules.

Table 1. Retrospective tests using SLASH fragments

Decile	R2	SAF	NSD	SAS
1st	26	26	25	20
2nd	18	15	8	8
3rd	9	8	11	10
4th	4	6	6	5
5th	3	2	1	2
6th	0	0	2	4
7th	0	2	2	4
8th	0	1	1	2
9th	0	0	3	2
10th	0	0	1	3
ES	5.03	6.69	10.5	14.7
PM	12.4	12.4	13.2	14.8

Table 2. Leave-one-out tests using SLASH fragments

Decile	R2	SAF	NSD	SAS
1st	14	12	17	11
2nd	18	18	9	13
3rd	6	7	4	4
4th	4	6	8	7
5th	3	5	6	8
6th	6	1	5	1
7th	3	4	4	5
8th	4	3	1	5
9th	1	4	3	3
10th	1	0	3	3
ES	13.7	14.6	16.4	19.1
PM	16.0	18.8	16.4	18.4

Table 3. Predictive test using SLASH fragments

Decile	R2	SAF	NSD	SAS
1st	6	6	5	3
2nd	3	3	2	5
3rd	4	2	4	3
4th	2	2	0	1
5th	0	2	2	1
6th	1	0	2	1
7th	1	2	2	3
8th	0	0	0	0
9th	0	0	0	0
10th	0	0	0	0
ES	2.64	3.12	3.55	3.96
PM	17.6	17.6	17.6	20.8

COMPARISON OF THE FOUR FRAGMENT WEIGHTS, USING REAL DATA

The fragments generated using SLASH were tested in two ways. First, they were used to calculate the fragment weights described above. These weights were then used to predict the activity of untested compounds in an attempt to establish how much insight the fragment weights were providing into the activities of the compounds in the training set. Second, the

Table 4. Number of fragments for large test 1

Level	Chains	Groups	Rings
0	1	1	1
1	22	22	36
2	179	266	1 144
3		787	2 054
4			2 493

fragments were used to form a bit string analogous to a Daylight fingerprint, to see whether the fragments could be used as the basis of a similarity measure. The investigations thus focused on two typical situations: (1) where several molecules have been tested and one wants to find substructural motifs that are associated, positively or negatively, with that activity, and (2) where there is just a single active compound, and one wants to find the nearest neighbors for biological testing.

Using SLASH fragments

Initial tests using a small data set Initially, three sets of tests were carried out on a small set of 500 compounds. These compounds had been tested in an HTS screen and 60 of them were active. In each of the tests, the compounds were divided into a training set and a test set and SLASH fragments were generated for both sets. The fragment weights were calculated for the training set, using the appropriate formula. These fragment weights were then used to calculate a relative probability of activity of the test compounds by calculating the average weight of the fragments each test molecule contained. This allowed the test molecules to be ranked in descending order of likely activity. In the ideal case, this ordering would result in all the active compounds being placed before all the inactive ones in the test set. Three measures were used to assess how far the results deviated from this ideal situation:

Percent misplaced (PM) This measure gives the percentage of compounds that were classified active (i.e., appeared in the top 60 places in the ranking) but were in fact inactive and vice versa.

Error score (ES) If there were 10 compounds in the test set, 4 of which were active, the ideal predicted ranking would be AAAAIIIIII. Consider the two rankings IAAAIIIIIA and AAAIAIIIII. These would both have a PM score of 20, but the second ranking, at a glance, appears to be much the better result, since the only error is that compounds 4 and 5 are swapped, whereas the first ranking has an active compound as the compound predicted least likely to be active.

Table 5. Results for large test 1

Decile	R2 Retrospective	R2 Leave-one-out
1st	212	139
2nd	44	43
3rd	23	13
4th	14	22
5th	10	13
6th	2	9
7th	0	19
8th	0	24
9th	0	12
10th	0	11
ES	26.3	81.8
PM	1.78	2.14

Table 6. Number of fragments for large test 2

Level	Chains	Groups	Rings
0	1	1	1
1	27	20	36
2	288	329	954
3		963	1 839
4			2 222

Table 7. Results for large test 2

Decile	R2 Retrospective	R2 Leave-one-out
1st	739	570
2nd	360	315
3rd	227	228
4th	210	187
5th	110	127
6th	69	70
7th	68	63
8th	38	76
9th	1	99
10th	0	87
ES	311	495
PM	8.79	9.92

Table 8. Results using R2 weights and daylight fingerprint fragments

Decile	Retrospective	Leave-one-out	Predictive
1st	27	18	5
2nd	20	12	3
3rd	10	4	1
4th	2	6	2
5th	0	5	1
6th	1	5	2
7th	0	4	1
8th	0	1	0
9th	0	4	2
10th	0	1	0
ES	5.94	17.1	4.87
PM	11.2	16.0	16.0

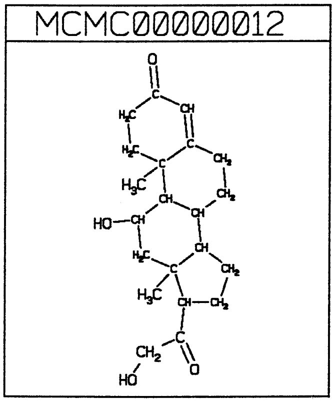

Figure 5. MCMC00000012.

The error score attempts to quantify this sort of distinction. It is calculated by marking the point in the list where one would expect the actives to stop and the inactives to start, (between 4 and 5 in the current example). The distance of each misclassified compound from this point is then summed, for both active and inactive compounds, and the result divided by the number of compounds in the set. This gives a score of 1.0 for the first ranking [(4 + 6)/10] and 0.2 for the second [(1 + 1)/10]. Thus a lower error score means a better ranking for that particular set of compounds.

Comparison of decile rankings The number of active compounds in each decile of the activity ranking was found and two rankings compared by the distribution of active compounds within the deciles.

The three different tests that were carried out were as follows: a retrospective prediction, wherein the same set of compounds formed both the training and test sets; a leave-one-out test, wherein each compound was predicted using the rest as the training set; and a full predictive test, wherein the 500 compounds were divided into two sets of 375 and 125 compounds, with the former being the training set. Owing to the random manner in which the compounds were split into the two sets, there were 43 active compounds in the training set, with 17 in the prediction set. In each case, a whole set of calculations was performed, using fragments at different levels. Only the results for all fragments at all levels are presented here, since the lower fragment levels proved to be insufficiently discriminating to be of much use.

These results are presented in Tables 1–3. It can be seen from Tables 1–3 that the fragments go some way in explaining the activity of an appreciable number of the compounds in each set. One would not expect complete accuracy of prediction. First, the fragment-based analysis treats the fragments in isolation, without reference to other fragments in the molecule. Second, the compound tested might not necessarily have the structure listed in the database. The structures were taken from the Zeneca Pharmaceuticals compound collection, which has been accumulated over several decades. It is inconceivable that in that time some of the samples have not deteriorated, and indeed there is evidence that this is in fact the case.

The Robertson and Sparck–Jones weighting formula can be

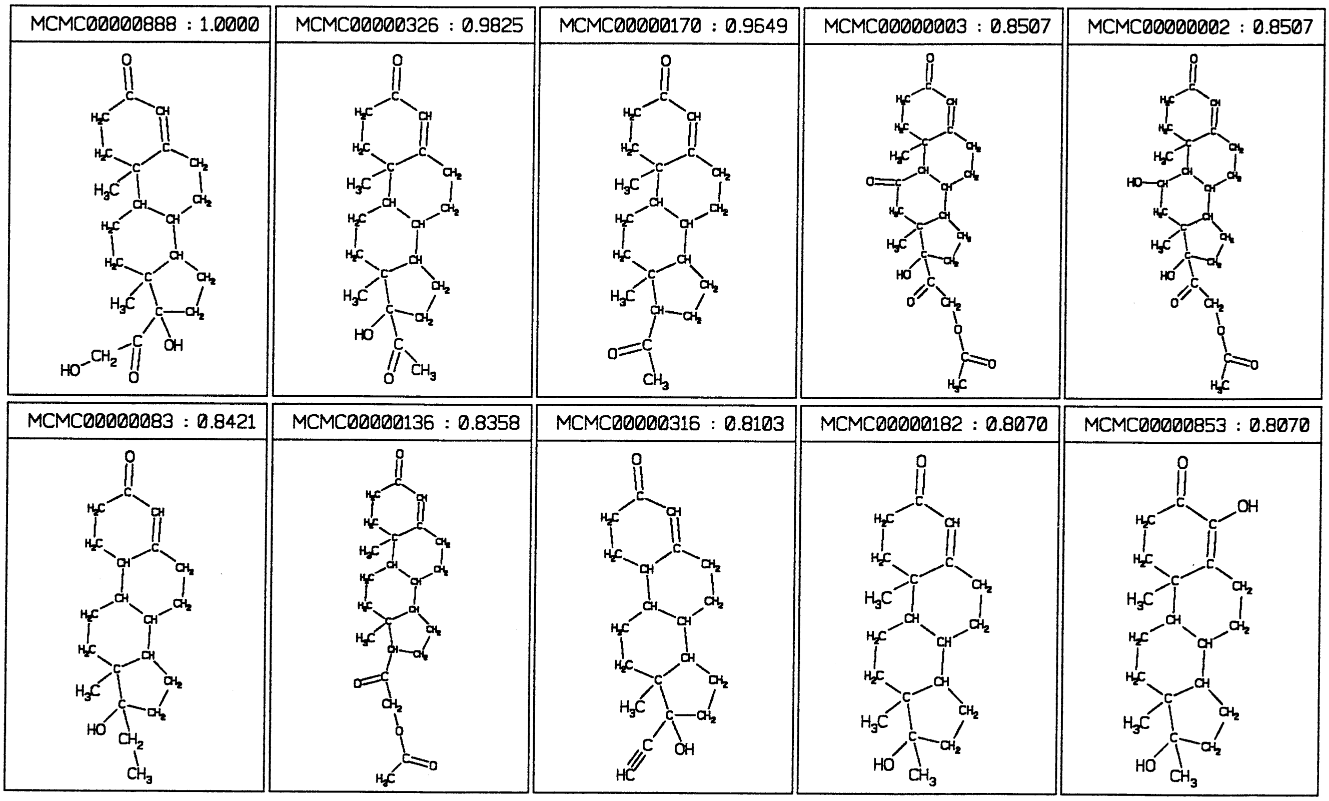

Figure 6. Molecules similar to MCMC00000012, using SLASH fragments.

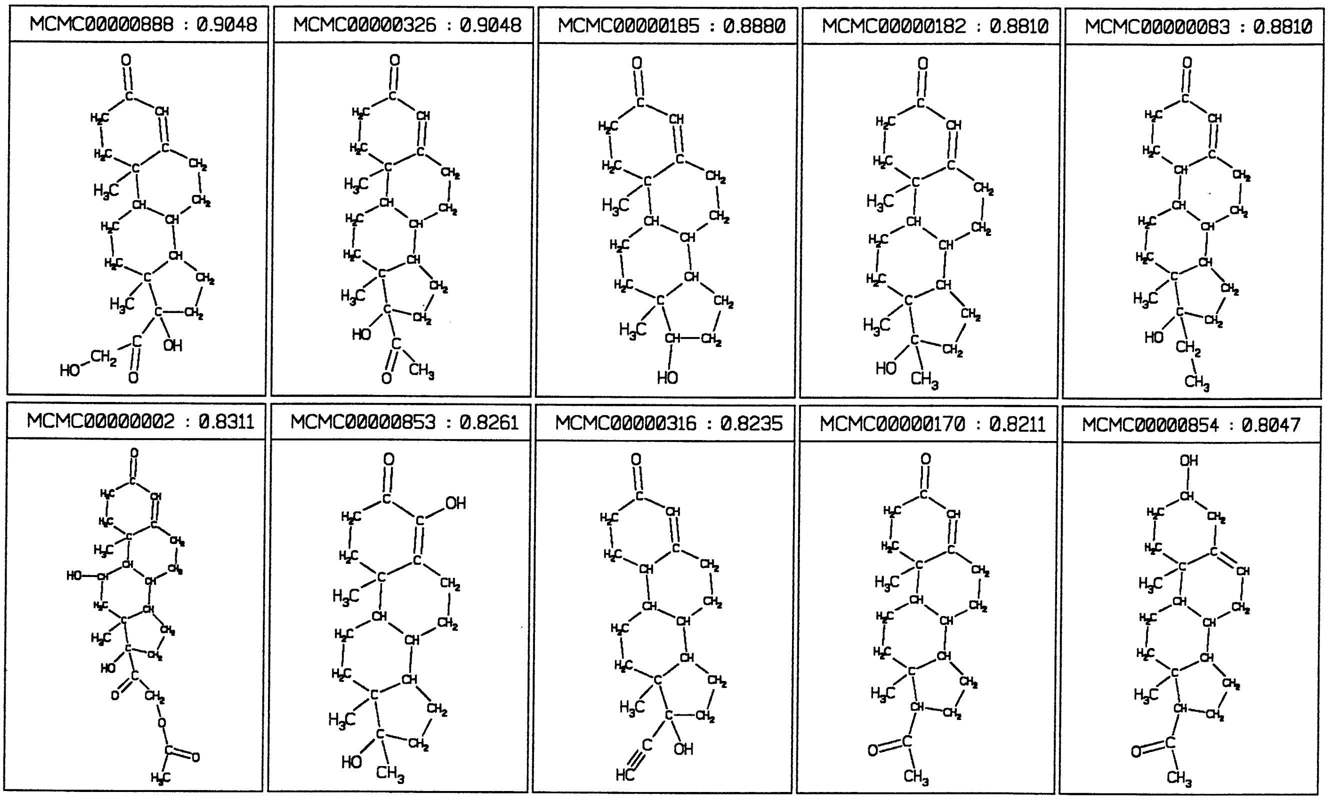

Figure 7. Molecules similar to MCMC00000012, using Daylight fingerprints.

seen to be consistently better than the other three in these tests, when looking at either the error score, the percentage misplaced, or the decile rankings. All four scoring methods have much better performance in the retrospective tests (Table 1) than in the predictive tests (Table 3). This is as expected, since the latter test is far more stringent.

Larger tests Having established the working principles using a small test set, two much larger tests were carried out, again using real HTS data. Owing to the much larger data sets, the tests could not be as detailed. Only the results for the R2 weighting formula are discussed here; other, inferior results have been described by Cosgrove.[21] Neither test set was chosen for any reason other than the ready availability of the data.

Results of test 1 The first test was on 25 000 molecules, of which 305 were active and the rest inactive. In total, 784 458 fragments were produced, an average of 31.4 per molecule. The breakdown of fragment types is given in Table 4, with the results of the fragment weighting tests in Table 5. The results of this are encouraging. In the retrospective case, more than two-thirds of the active compounds are in the first decile and all bar two of the compounds are in the top half of the ranking. The leave-one-out rankings are less impressive, but there are still one-third of the active compounds in the first decile, with 83% in the top half of the ranking. The difference is relatively easy to explain. The R2 score is particularly prone to large values when there is only one fragment of a particular type. With such a small proportion of active compounds in this set of molecules, there are a number of such singleton fragments. In the retrospective case, this will result in the whole compound receiving a high score—in effect, the fragment is automatically flagging the compound as active. With the leave-one-out calculation, this effect necessarily vanishes, resulting in a marked deterioration of the results.

Results of test 2 The second test set contained 28 456 molecules, of which 1 822 were active. These molecules produced 979 256 fragments, an average of 34.4 per molecule. The fragment breakdown is given in Table 6, with the fragment weight results in Table 7. The difference between the retrospective and leave-one-out results is rather less in this case. This may be due to the larger proportion of active compounds, as discussed above. The results are, on the whole, poorer than those of the first large test set, but there are still 78% of the active compounds in the upper half of the ranking.

Using Daylight fragments

As a comparison, the same calculations were carried out on the small data set, using Daylight "fragments." These fragments were approximated by using the individual bits of the fingerprint for each molecule to denote a fragment. This is not entirely comparable to the SLASH case, as each fragment may set more than 1 bit, and some fragments may set the same bit. The former is not a large problem, since it merely duplicates information, and whilst the latter could give rise to noise in the data, the final bit density was only 12.3%, so this should not be a major concern. The fingerprints were created using version 4.34 of the Daylight SMILES and fingerprint toolkits, using 1 024 bits and no folding. Table 8 shows the results for retro-

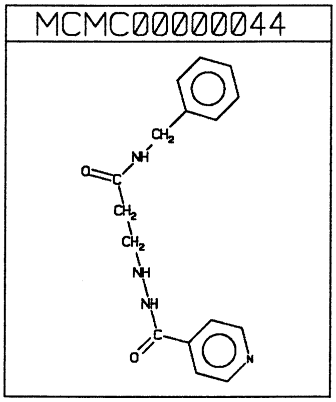

Figure 8. MCMC00000044.

spective, leave-one-out and predictive tests using the R2 weight. Cosgrove[21] reports other, poorer results using the other three weights. Comparison with Tables 1–3 shows that the SLASH results are uniformly but not overwhelmingly better. The results for the other weights are also poorer for the Daylight fragments than for the SLASH fragments.[22]

SLASH FRAGMENTS AND MOLECULE SIMILARITIES

It is possible to form, using the SLASH fragments, a bit string that allows the calculation of similarity between molecules using, for instance, the Tanimoto coefficient of similarity.[22] The Daylight fingerprint is composed by forming all possible atom paths up to a given number of atoms, with the presence of a fragment being flagged by the setting of 3 or 4 bits in the fingerprint. Each fragment does not set a unique bit, but it is likely that the vast majority of fragments will set a unique pattern of this small number of bits, thus enabling the effect of the fragment to be distinguished in the bit string. This procedure is necessitated by the large number of possible fragments. There are far fewer fragments produced by the SLASH analysis; for instance, with a molecule set of 25 000 compounds, about 5 000 different fragments are produced. This makes it possible to associate a particular fragment with its own bit in a fingerprint. The advantage of this method is that it is then relatively easy to establish exactly what fragments two molecules have in common that are giving rise to the measured similarity, something that is not possible with the Daylight fingerprints. The hierarchical nature of the SLASH fragments enables both general and specific similarities to be measured, since, for instance, all compounds containing a five-membered aromatic heterocycle will set the same bit, even if the heteroatoms are different in different rings, a difference that will be apparent by the setting of bits for the higher level ring fragments. The user can control what levels of fragment go into the bit string, so that this generality can be dropped if necessary.

As well as bits for each fragment, the SLASH fingerprint also has bits set aside for fragment connection paths. A connection table is created for the level 0 fragments in each molecule, giving a crude picture of how the fragments are joined together. From this connection table, all possible frag-

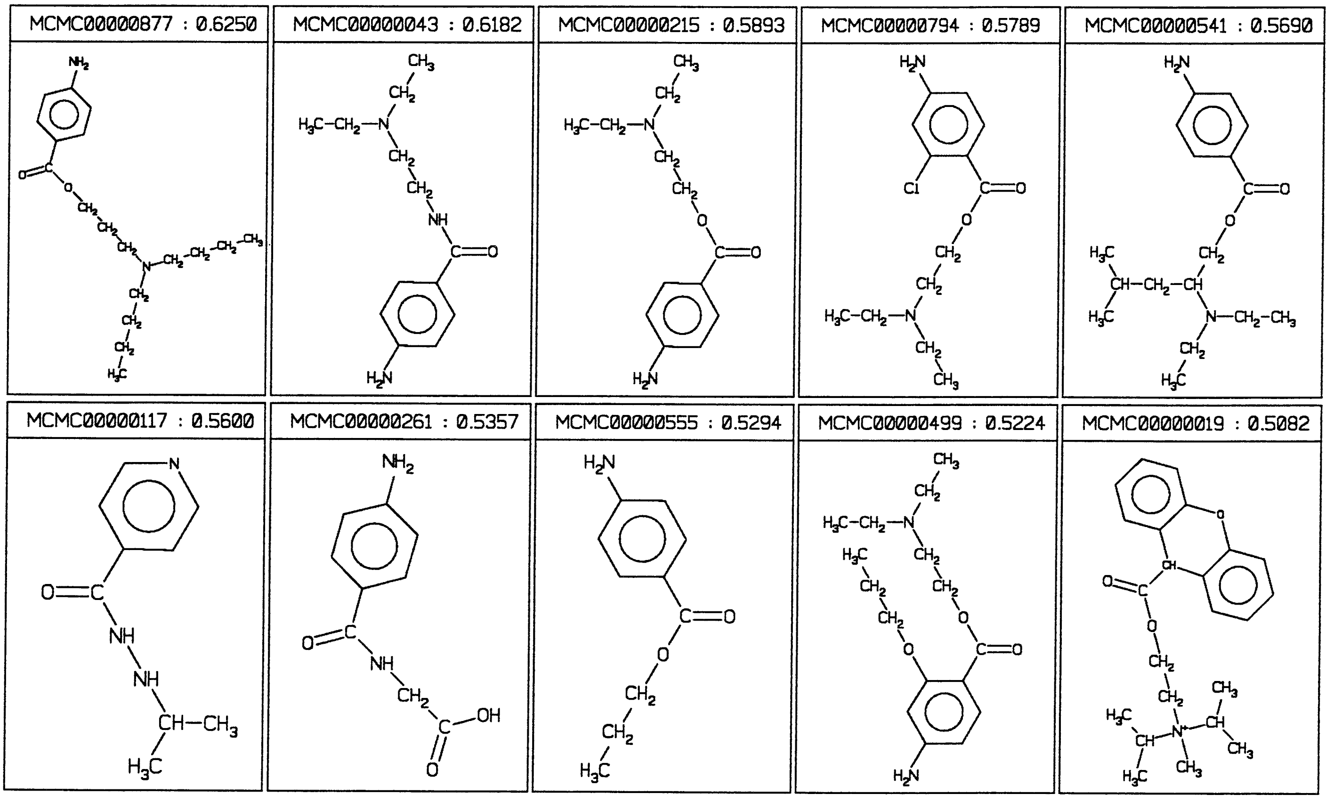

Figure 9. Molecules similar to MCMC00000044, using SLASH fragments.

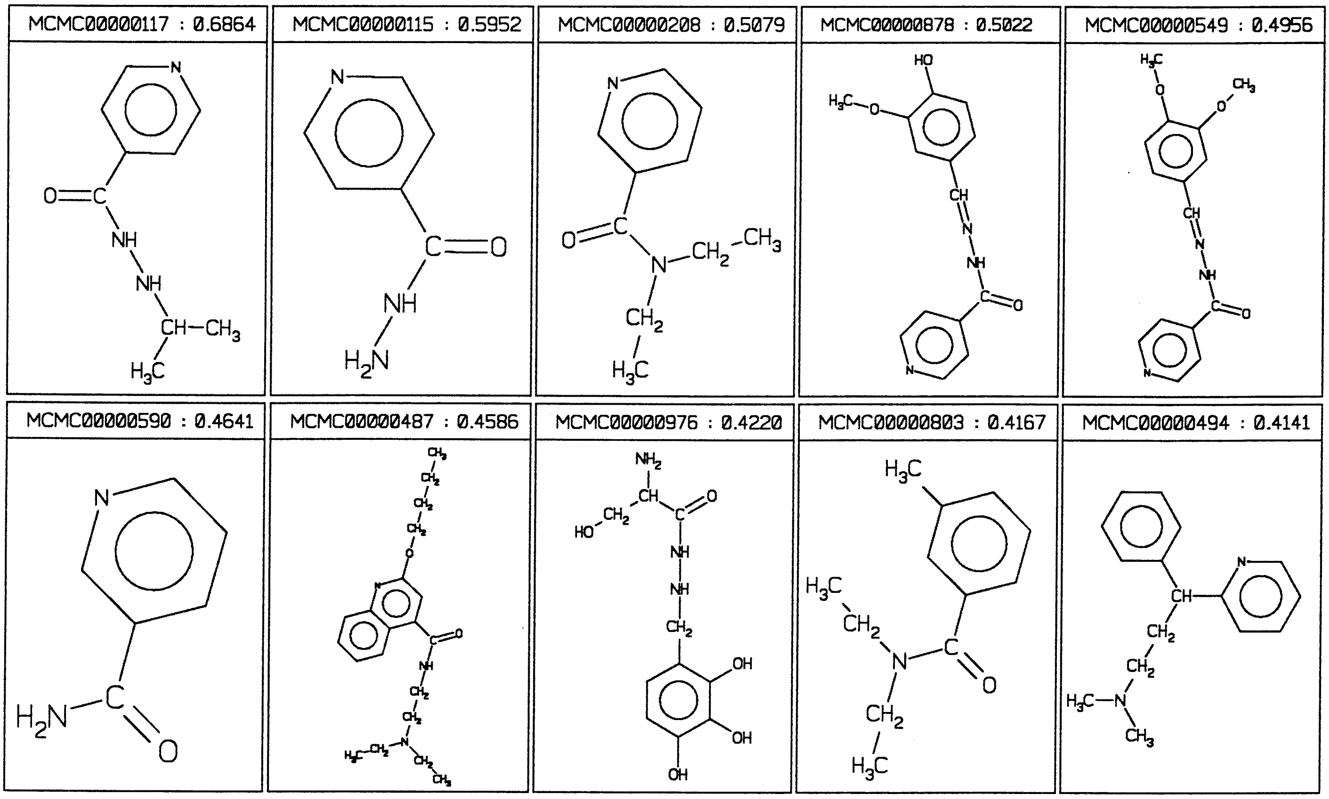

Figure 10. Molecules similar to MCMC00000044, using Daylight fingerprints.

ment paths are enumerated in a manner directly analogous to the formation of Daylight fingerprints. However, unlike the Daylight fingerprints, it is possible to allocate 1 bit per connection path, because there are again far fewer possibilities. This is because there are only three different entities at level 0 (ring, group, or chain), and the paths are all quite short. The inclusion of these fragment paths in the fingerprint enables the distinguishing of molecules that contain the same fragments, but have different structures.

Obviously, the assessment of a measure of chemical similarity is a largely subjective process, since different chemists with different biases attach different levels of significance to various features. One slightly more objective test would be to take a set of compounds and, for each active compound, rank the rest of the compounds in order of descending similarity. If the activities of the compounds are due to a specific effect, one would expect most of the active compounds to appear at the top of the similarity ranking, with the inactive ones at the end. This has been done for each of the active compounds in the smaller, 500-compound test set used above, with 60 actives, and the number of actives in each of the first 60 positions in the ranking list computed. In the ideal case, with the 60 actives coming first, followed by the 440 inactives, this would give a value of 60.0. In the case of the actives being randomly distributed through each ranking list, the value would be 7.2 [(60 × 60)/500]. The values for the SLASH and Daylight similarity methods were 9.88 and 11.29, respectively. Neither of these values is especially impressive, but the Daylight similarity is clearly slightly better.

A second test, which is much more subjective, is to take a set of compounds, pick a sample few, and examine the 20 most similar to it, found with the two similarity measures. One can then compare the two lists, to see which appears more sensible according to one's own personal bias. This has been done for 992 compounds taken from the MACCS Comprehensive Medicinal Chemistry (CMC) database. Some examples of similar compounds are shown in Figures 5–13. Figure 5 shows MCMC00000012, with Figures 6 and 7 giving the 10 compounds most similar to it, using SLASH and Daylight fingerprints, respectively. The actual similarity values are not directly comparable, but the orderings should be. Both fingerprints find a number of compounds highly similar to the target compound, and there is a large overlap between the two sets, with 8 of the

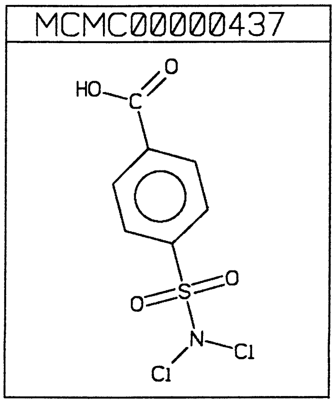

Figure 11. MCMC00000437.

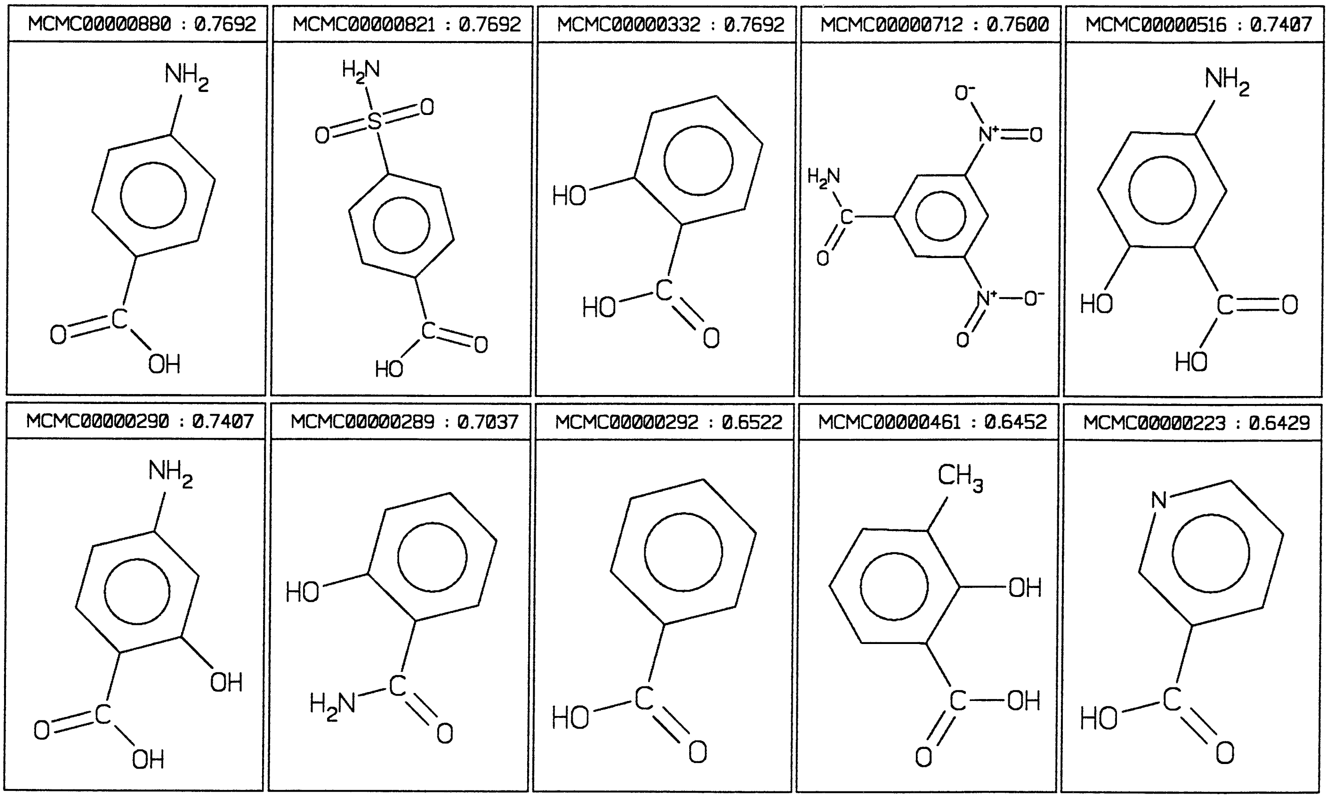

Figure 12. Molecules similar to MCMC00000437, using SLASH fragments.

10 compounds being the same in both sets. Figures 8–10 show MCMC00000044 and its 10 most similar compounds, using SLASH and Daylight. The first thing to note is that the similarity values in this instance are much lower in all instances than for the previous set of molecules. There are no compounds in common, although 6 of the first 20 are the same. For instance, MCMC00000208 is third similar in the Daylight ranking, but only twelfth in the SLASH list. The other compounds in the two lists show some interesting differences. The target compound is an aryl amide, as are all the Daylight compounds that are also carboxylates. SLASH, on the other hand, also finds a selection of esters that are far more similar in structure to the target compound, ignoring the amide/ester change, than are the nonamide compounds that the Daylight fingerprints give as similar. Whether this is desirable or not is a matter of opinion, but if one were looking for new synthetic targets that are similar, but not too similar, to a given compound, the SLASH list might be preferable. This effect is emphasised in Figures 11–13, which are the analogous results for MCMC00000437. In this case, the molecules found using the Daylight fingerprints all contain a sulphur atom. The SLASH fingerprints, on the other hand, pull out compounds that are markedly different, with only one containing a sulphur. They do, however, contain far more carboxylic acids, a group that is also found in the target compound. In this case, one's opinion of which set of compounds is more similar would be coloured by the sort of chemistry one was interested in performing.

The reason for the differences in the two similarity measures is the hierarchical nature of the SLASH fragments. The esters that the SLASH fingerprints perceived to be similar to MCMC00000044 are found because the ester group has the same number of noncarbon atoms in the same connection pattern as the amide group. Thus, at the more general levels of group hierarchy, the two groups are considered equivalent, and will in fact set 3 bits in the fingerprint that are the same, and only 1 bit that is different.

CONCLUSIONS

This article has described the program SLASH, which gives a novel way of looking at active and inactive molecules in terms of the functional groups they contain. The fragments generated by SLASH have been shown to be of some limited use in predicting the activities of untested compounds based on the knowledge of the structures of tested ones. From this one might tentatively suggest that the fragments might be of some use in establishing the functional groups that are responsible for the active compounds in a set, with a view to forming a pharmacophoric hypothesis for a particular biological test. One potential use might be to take a set of active structures, which might have been clustered by some other method, and try to establish which functional groups they have in common that also have high fragment scoring weights, indicating an association with activity. This would help prevent the generation of spurious pharmacophoric suggestions based on functional groups that are present in large numbers in the active compounds, but that are also present in large numbers in the inactive compounds. Such coincidences might not be spotted unless the chemist examining the active compounds had considerable knowledge

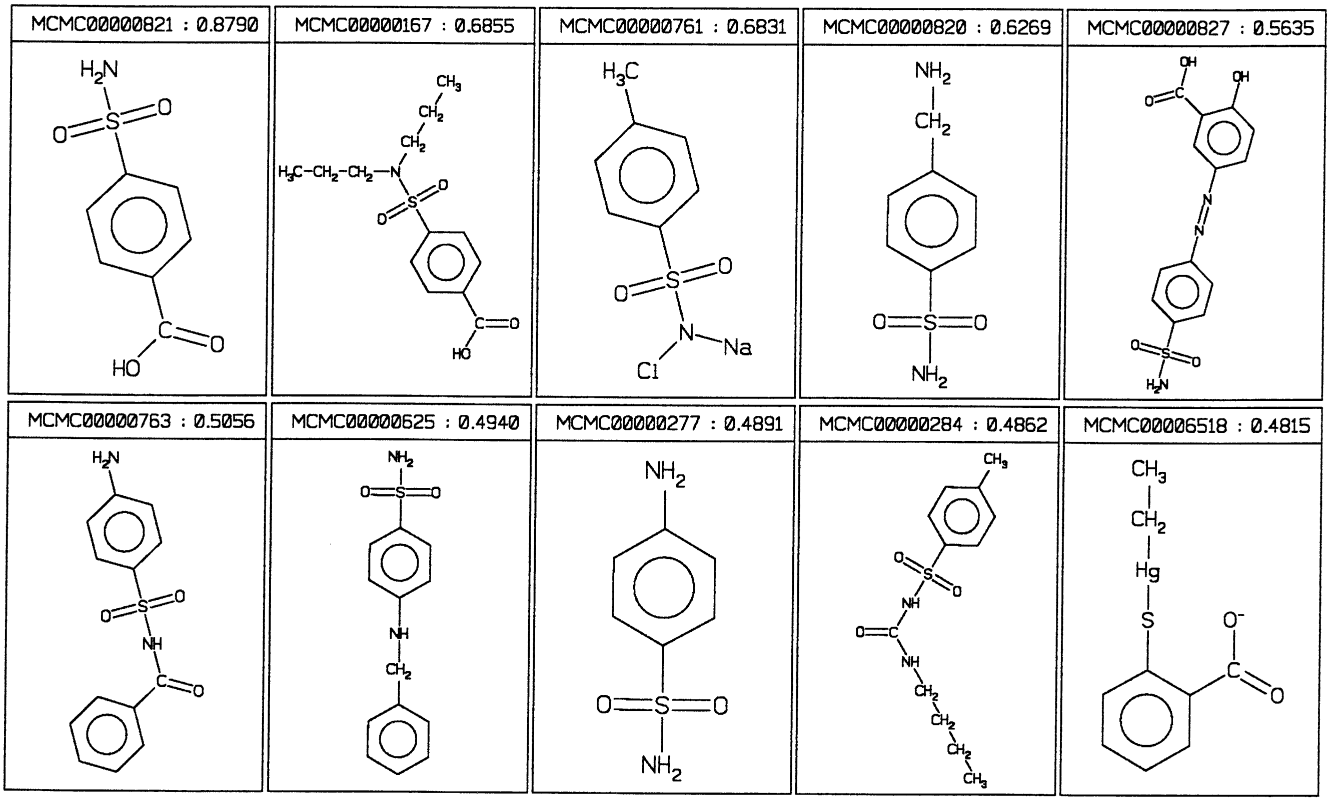

Figure 13. Molecules similar to MCMC00000437, using Daylight fingerprints.

of the spread of many functional groups throughout the whole set of molecules that were tested.

The fragments generated by SLASH have also been investigated as the basis for a measure of molecular similarity. They are less successful in this area, and would not be preferred over, for instance, Daylight fingerprints. However, there might be a use for them in suggesting new possibilities for synthesis to take a project outside a well-investigated area of chemical space.

ACKNOWLEDGMENTS

D.A.C. acknowledges Zeneca Pharmaceuticals for funding for this study.

REFERENCES

1 Berkoff, C.E., Cramer, R.D., III, and Redl, G. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* 1974, **17,** 533–535

2 Geran, R.I., Hazard, G.F., Hodes, L., and Richman, S. A statistical–heuristic method for the automated selection of drugs for screening. *J. Med. Chem.* 1977, **20,** 469–475

3 Hodes, L. Selection of molecular fragment features for structure–activity studies in anti-tumour screening. *J. Chem. Inf. Comput. Sci.* 1981, **21,** 132–136

4 Adamson, G.W. and Bush, J.A. Method for relating the structure and properties of chemical compounds. *Nature (London)* 1974, **248,** 406–407

5 Adamson, G.W. and Bush, J.A. Evaluation of an empirical structure–activity relationship for property prediction in a structurally diverse group of general anaesthetics. *J. Chem. Soc. Perkin I* 1976, 168–172

6 Carhart, R.E., Smith, D.H., and Venkataraghavan, R. Atom pairs as molecular features in structure–activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* 1985, **25,** 64–73

7 Deleted in proof.

8 Adamson, G.W. and Bawden, D. A method of structure–activity correlation using Wiswesser line notation. *J. Chem. Inf. Comput. Sci.* 1975, **15,** 215–220

9 Adamson, G.W. and Bawden, D. Substructural analysis techniques for empirical structure–property correlation. Application to stereochemically related molecular properties. *J. Chem. Inf. Comput. Sci.* 1980, **20,** 97–100

10 Adamson, G.W. and Bawden, D. An empirical method of structure–activity correlation for polysubstituted cyclic compounds using Wiswesser line notation. *J. Chem. Inf. Comput. Sci.* 1976, **16,** 161–165

11 Adamson, G.W. and Bawden, D. A substructural analysis method for structure–activity correlation of heterocyclic compounds using Wiswesser line notation. *J. Chem. Inf. Comput. Sci.* 1977, **17,** 164–171

12 Sasaki, S., Sukekawa, M., and Takahashi, Y. Automatic identification of molecular similarity using reduced-graph representation of chemical structures. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 639–643

13 Brugger, W.E., Jurs, P.C., and Stuper, A.J. Generation of descriptors from molecular structures. *J. Chem. Inf. Comput. Sci.* 1976, **16,** 105–110

14 Chou, J.T., Jurs, P.C., and Yuan, M. Computer-assisted structure–activity studies of chemical carcinogens. A heterogeneous data set. *J. Med. Chem.* 1979, **22,** 476–483

15 Blankley, C.J., Humblet, C., Shemetulskis, N.E., Weininger, D., and Yang, J.J. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 862–871

16 Ormerod, A., Willett, P., and Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct. Activity Relat.* 1989, **8,** 115–129

17 Robertson, S.E. and Sparck-Jones, K. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* 1976, **27,** 129–146

18 Klopman, G. and Rosenkranz, H.S. Toxicity evaluation by chemical substructural analysis—the Tox-II program. *Toxicol. Lett.* 1995, **79,** 145–155

19 Klopman, G. and Macina, O.T. Computer-automated structure evaluation of antileukemic 9-aniloacridines. *Mol. Pharmacol.* 1986, **136,** 67–77

20 Ormerod, A., Willett, P., and Bawden, D. Further comparative studies of Fragment weighting schemes for substructural analysis. *J. Am. Chem. Soc.* 1984, **106,** 7315–7321

21 Cosgrove, D.A. The Automated Search for Chemically Interesting Fragments in Molecular Databases. M. Phil. thesis. University of Sheffield, Sheffield, England, 1996

22 Willett, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press, Letchworth, UK, 1987, p. 54