ELSEVIER

Comparison of algorithms for dissimilarity-based compound selection

Michael Snarey,* Nicholas K. Terrett,* Peter Willett,† and David J. Wilton†

**Pfizer Central Research, Sandwich, Kent, United Kingdom*
†Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield, United Kingdom

Dissimilarity-based compound selection has been suggested as an effective method for selecting structurally diverse subsets of chemical databases. This article reports a comparison of several maximum-dissimilarity and sphere-exclusion algorithms for dissimilarity-based selection. The effectiveness of the algorithms is quantified by the numbers of biological activity classes identified in subsets selected from the World Drugs Index database, and by the numbers of active compounds identified in feedback searches of this database. The experiments demonstrate the general effectiveness and efficiency of the MaxMin algorithm. © 1998 by Elsevier Science Inc.

INTRODUCTION

The requirements of high-throughput screening and combinatorial synthesis programmes have led to much interest in the development of computer-based methods for selecting sets of structurally diverse compounds from chemical databases.[1] Here, we focus upon methods for *dissimilarity-based compound selection* (DBCS),[2] which involves the identification of a subset comprising the n most dissimilar molecules in a database containing N molecules (where, typically, $n \ll N$). The identification of the optimally diverse subset (using some quantitative index of diversity) is computationally infeasible, since it requires consideration of all possible n-member subsets of the database, unless both n and N are small. Practicable approaches to DBCS hence involve approximate procedures that are not guaranteed to result in the identification of the most dissimilar possible subset;[2–13] that said, there is evidence to suggest that the subsets identified are only marginally suboptimal.[14]

Several different DBCS algorithms have been reported in the literature and it is natural to ask which is the most appropriate for use in lead-discovery programmes, in much the same way as several authors have sought to establish the most appropriate descriptor types for diversity analysis.[15–18] The basic rationale underlying automated compound-selection algorithms is the assumption that a set of diverse compounds that spans structural space (however that is defined, e.g., in terms of physicochemical parameters, substructural fragments, or topological indices) will also span biological activity space.[19] It should thus be possible to compare the effectiveness of different structure-based selection methods by the extent to which they result in sets of compounds that exhibit a wide range of biological activities.[18] This idea forms the basis for the work reported here, which provides a quantitative comparison of the effectiveness of different DBCS algorithms; in fact, we consider two subclasses of this class of selection algorithm, which we shall refer to as *maximum-dissimilarity* algorithms and *sphere-exclusion* algorithms. We compare them both with each other and with random selection, since it has been suggested that computer-based procedures are no better than random at selecting bioactive molecules.[20–22]

Address reprint requests to: P.WILLETT@SHEFFIELD.AC.UK.

Received 1 December 1997; revised 24 February 1998; accepted 2 March 1998.

THE ALGORITHMS

Maximum-dissimilarity algorithms

A simple algorithm for selecting a *Subset* of size n (i.e., containing n molecules) from a *Database* of size N is shown in Figure 1. This algorithm, which was first described by Kennard and Stone[23] almost three decades ago, permits many variants depending upon the precise implementation of Steps 1 and 3.

Possible mechanisms for the choice of the initial compound in Step 1 include choosing a compound at random, choosing that compound that is most dissimilar from the other compounds in *Database*, or choosing that compound that is nearest to the centre (in some sense) of *Database*, relative to the others. In our experiments we initialised *Subset* with that molecule that has the largest sum of dissimilarities (smallest sum of similarities) relative to the other molecules in *Database*.

Step 3 in Figure 1 requires a quantitative definition of the dissimilarity between a single compound in *Database* and the

Journal of Molecular Graphics and Modelling 15, 372–385, 1997

655 Avenue of the Americas, New York, NY 10010
1093-3263/97/$17.00
PII S1093-3263(98)00008-4

1. Initialise *Subset* by transferring to it a compound from *Database*.
2. Calculate the dissimilarity between each remaining compound in *Database* and the compounds in *Subset*.
3. Select that compound from *Database* that is most dissimilar to *Subset* and transfer it to *Subset*.
4. Return to Step 2 if there are less than *n* compounds in *Subset*.

Figure 1. Maximum-dissimilarity algorithm.

group of compounds that comprise *Subset*, so that the most dissimilar molecule can be identified in each iteration of the algorithm. There are several ways in which "most dissimilar" can be defined[11], with each definition resulting in a different version of the algorithm. Here, we have used two such definitions, which we refer to as *MaxSum* and *MaxMin*. Let *DIS*(*A*, *B*) be the dissimilarity between two molecules, or sets of molecules, *A* and *B*. Consider a single compound, *J*, taken from *Database* and the *m* compounds that form the current membership of *Subset* at some stage in the selection process; the dissimilarity between *J* and *Subset* is then given by

$$DIS(J, Subset) = \sum DIS(J, K)$$

and

$$DIS(J, Subset) = minimum\{DIS(J, K)\}$$

in the case of the MaxSum and MaxMin definitions, respectively, with $K(1 \leq K \leq m)$ ranging over all of the *m* molecules in *Subset* at that point. The molecule chosen for addition to *Subset* is then that with the largest value of *DIS*(*J*, *Subset*).

We have chosen these two definitions from amongst the many possible for use in Step 3 since both have been discussed previously in the literature and since both can be implemented to allow the rapid processing of large databases. The general maximum-dissimilarity algorithm outlined in Figure 1 has an expected time complexity of $O(n^2N)$ for the selection of an *n*-member *Subset* from an *N*-member *Database*; as *n* is typically some nontrivial fraction of *N* (for example, Brown and Martin[17] suggest that *n* should be about $0.2N$ if a subset is to be fully representative of its parent database, while Matter[18] suggests a value of about $0.35N$), this corresponds to an expected time complexity of no less than $O(N^3)$. It is, however, possible to find more efficient algorithms for specific definitions of dissimilarity, and we have used implementations of the $O(nN)$ algorithms for the MaxSum and MaxMin definitions that have been described by Holliday et al.[10] and by Polinsky et al.,[13] respectively.

In addition to the basic algorithm shown in Figure 1, we also test a minor modification in which the compound selected in Step 2 is checked to ensure that it is not too similar to one that has already been selected. This is done by setting a threshold dissimilarity, *t*, and then rejecting the selected compound if it has a dissimilarity less than *t* with any of the compounds already in *Subset*. We have adopted the figure of 0.15 for the complement of the Tanimoto similarity, as advocated by Matter[18] and by Ferguson et al.[19]

Sphere-exclusion algorithms

Here, a dissimilarity threshold *t* is set, which can be thought of as the radius of a hypersphere in multidimensional chemistry space. The basic algorithm described by Hudson et al.[12] proceeds by selecting a compound at each stage and then excluding from further consideration all those other compounds within the sphere centred on that compound, as shown in Figure 2.

Many variants of this general algorithm are possible depending upon the selection criterion that is adopted in Step 2 and upon the value that is chosen for the threshold similarity, *t*. Hudson et al. initialise *Subset* by choosing that compound that has the smallest sum of dissimilarities relative to the rest of *Database*, and in subsequent iterations choose that compound from *Database* that is least dissimilar to *Subset*. This suggests at least three "least dissimilar" variants: SE-MinSum, in which the compound with the smallest sum of dissimilarities is selected each time; SE-MinMin, in which the compound with the

1. Define a threshold dissimilarity, *t*.
2. Select a compound, *J*, from *Database* and transfer it to *Subset*.
3. Remove from *Database* all compounds that have a dissimilarity with *J* of less than *t*.
4. Return to Step 2 if there are compounds remaining in *Database*.

Figure 2. Sphere-exclusion algorithm.

smallest minimum dissimilarity is selected; and SE-MinMax, the version advocated by Hudson et al., in which the compound with the smallest maximum dissimilarity is selected. In this last variant, the dissimilarity *DIS*(*J*, *Subset*) between some molecule *J* from *Database* and *Subset* is given by

$$DIS(J, Subset) = maximum\{DIS(J, K)\}$$

(cf. the MaxSum and MaxMin definitions given earlier). All of the sphere-exclusion algorithms studied here used the complement of the Tanimoto coefficient for calculating the *DIS*(*J*, *Subset*) values.

The various sphere-exclusion algorithms all start with the same compound, near the "centre" of the data set, and progressively work their way "outwards," little by little, until *Database* is exhausted. This is in marked contrast to the maximum-dissimilarity algorithms, in which the most dissimilar criterion in Step 3 of Figure 1 means that markedly disparate structures are selected right from the start of the processing. It would be possible to specify a maximum number of compounds for inclusion in *Subset*, and then to allow the sphere-exclusion algorithm to proceed until this number had been achieved (in a manner analogous to the maximum-dissimilarity algorithms). However, this might lead to an uneven sampling of *Database* since compounds towards its outer reaches could not be selected. Accordingly, our experiments have used the termination condition shown in Figure 2, with the result that it is not possible to specify, *a priori*, the precise final size of *Subset.*

A variant of the basic sphere-exclusion algorithm, referred to here as MDISS, is included in the DiverseSolutions package developed by Pearlman and co-workers at the University of Texas, in which the compound *J* that is selected in Step 2 of Figure 2 is chosen at random.[24] Unlike the previous sphere-exclusion algorithms, for which a given radius will always lead to the same subset being selected, the random element here means that a different subset will be selected each time that MDISS is run. The quality and size of subsets produced by MDISS is thus expected to be more variable than for sphere-exclusion algorithms that involve a nonrandom selection step.

It is difficult to estimate the time complexity of the sphere-exclusion algorithms. If the final subset contains *n* molecules then an inspection of Figure 2 shows that there are *n* iterations of the algorithm, each of which requires a selection step and then a similarity calculation involving all of the molecules remaining in *Database*. In the first iteration there will be $N - 1$ such molecules, giving a worst-case time complexity of $O(nN)$. However, the numbers of similarity calculations that need to be considered in each iteration will decrease as molecules are removed by application of the threshold dissimilarity (in Step 3 of the algorithm). The overall running time will hence be determined by the extent to which the number of *Database* compounds left after each iteration is less than the number after the previous iteration. MDISS is expected to be far more rapid than the other sphere-exclusion algorithms since it uses simple random selection in Step 2 of Figure 2, rather than a deterministic, dissimilarity-based selection procedure.

EXPERIMENTAL DETAILS

Our experiments have involved a subset of the compounds in the *World Drugs Index* (*WDI*), which contains many thousands of compounds that exhibit some kind of biological activity. The subset used here was obtained by excluding the following classes of molecule: those that are associated with nondrug activities, such as pesticides; those that are associated with very broad druglike activities, such as 'antiseptics'; those that contain elements other than C, H, N, O, S, and Hal; those that do not have a United States-approved name or an international nonproprietary name; those that do not give a valid CLOGP value; and those that are associated with more than one activity class. Removal of these various classes resulted in a file of 2049 compounds that belonged to 123 different activity classes. The compounds are not evenly distributed between the activity classes, with many of the activities being represented by just a single compound, as illustrated in Figure 3.

The molecules were characterised in one of two ways: by UNITY fingerprints, these being 988-member bit-strings describing the presence of various types of 2D substructural fragment[25]; and by 128 topological indices and physical properties generated by the Molconn-Z program.[26] Given these characterisations, the dissimilarity between a pair of molecules was calculated using the complement of either the cosine or the Tanimoto coefficient.[27] Note that the use of the fast centroid procedure of Holliday et al.[10] for the implementation of the MaxSum and SE-MinSum algorithms requires that the cosine coefficient be used; the other algorithms are not so restricted.

The various DBCS algorithms were used to select subsets of the 2049-molecule data set. As noted previously, selection effectiveness was measured by the number of different activity classes that were identified in a subset once it had been generated. This is an appropriate criterion for evaluating selection methods in the context of a typical random screening environment, where one wishes to identify a representative subset of a database that can then be submitted to a battery of bioassays. A more focused approach is appropriate if one is interested in a specific biological target, where the need is to identify compounds that are similar to ones that have already been shown to be active.

The similar property principle[28] suggests that such highly similar compounds, the *near neighbours*, have a high *a priori* probability of being active and thus are strong candidates for biological testing. A possible testing strategy that uses this *feedback* approach is outlined in Figure 4: the importance of feedback information in compound-selection algorithms has been noted previously by Taylor.[20] An effective selection algorithm will be one that chooses that *Subset* that maximises the number of actives identified in the second round of testing, and this number thus provides an alternative way of comparing the performance of the various DBCS algorithms. The near neighbours in Step 3 of Figure 4 can be identified using any of the many similarity measures that are now available.[28–30] Here, we specify a threshold value for a bit-string-based Tanimoto dissimilarity, *t*, and select as the near neighbours for biological testing all molecules from *Database* that have dissimilarities less than this threshold (in a manner that can be regarded as the inverse of the selection criterion used in the sphere-exclusion algorithms). The performance measure for each selection algorithm is then taken to be the numbers of actives and inactives from *Database* that are identified in Step 4 of the feedback search strategy outlined in Figure 4. The experiments here involved the eight activity classes in the *WDI* data set that contained the greatest numbers of molecules, namely 199 molecules classed as antibiotics, 115 as analgesics, 112 as cytostatics, 93 as corticosteroids, 78 as antiarteriosclerotics, 68 as

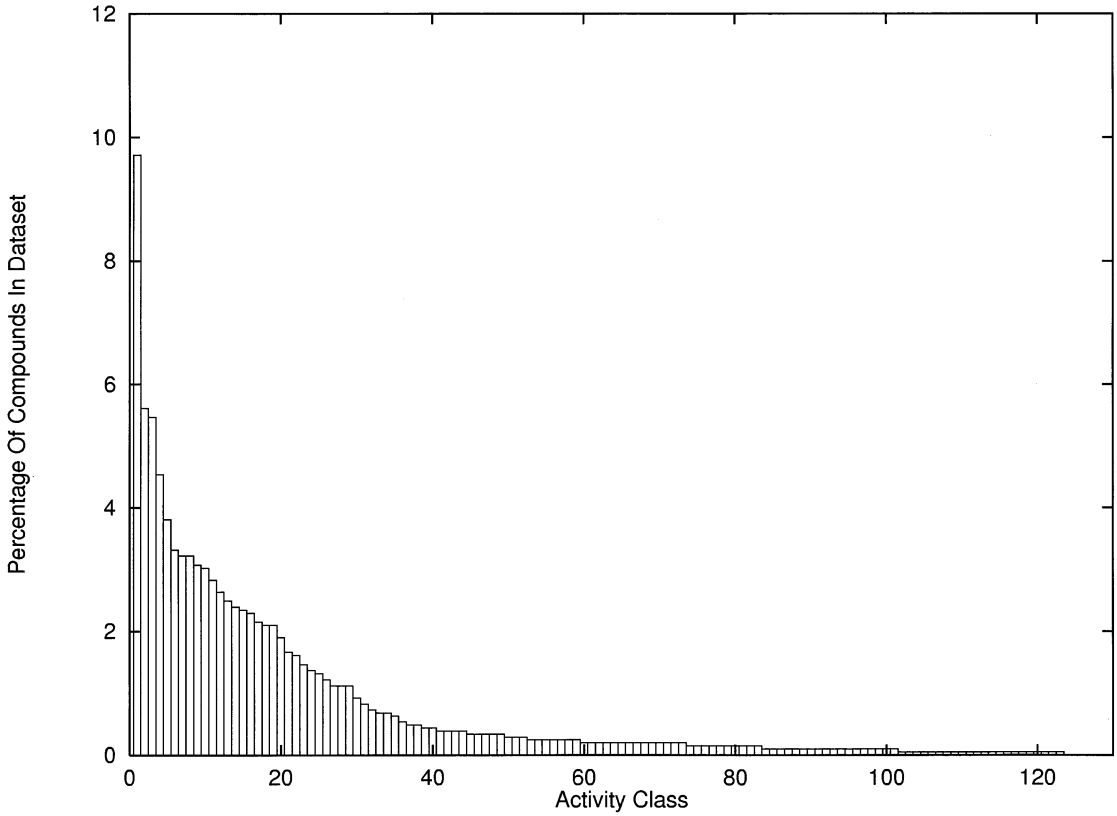

Figure 3. Frequencies of occurrence of the 123 activity classes in the 2 049-molecule WDI subset.

anthelmintics, 66 as spasmolytics, and 66 as H_1-antihistamines. The molecules were characterised by UNITY fingerprints, subsets chosen using the various algorithms described previously, and the near neighbours identified for the active molecules in each such subset.

RESULTS AND DISCUSSION

Figure 5 shows the numbers of different activity classes selected using the MaxSum (cosine coefficient) and MaxMin (both cosine and Tanimoto coefficients) algorithms with the UNITY data, for variously sized subsets. We also show the numbers that would be obtained using random selection, where the results are expressed by the mean values and 2-standard deviation error bars when averaged over 10 000 runs for each subset size; and using both optimal and pessimal selection strategies. The optimal results, using the chosen performance measure, are obtained with a selection algorithm that selects a member of a previously unrepresented activity class in each of its first 123 iterations, thereafter making up the remainder of the subset with additional compounds from previously selected activity classes. The pessimal results, using the chosen performance measure, are expected to be obtained with a selection algorithm that uses a MinMin criterion, i.e., an algorithm that selects at each step that molecule that is least (rather than most) dissimilar to the molecules already in *Subset*. This pessimal algorithm took the same starting molecule as for the other DBCS algorithms.

For subsets containing up to about 100 compounds, all the

1. Use a selection procedure to identify a diverse set of compounds from *Database*.
2. Test each of the compounds in *Subset* and identify those that are active in the assay of interest.
3. For each active compound identify those molecules remaining in *Database* that are sufficiently similar to suggest that they may also be active.
4. Test these near neighbours of the actives in *Subset*.

Figure 4. A testing strategy.

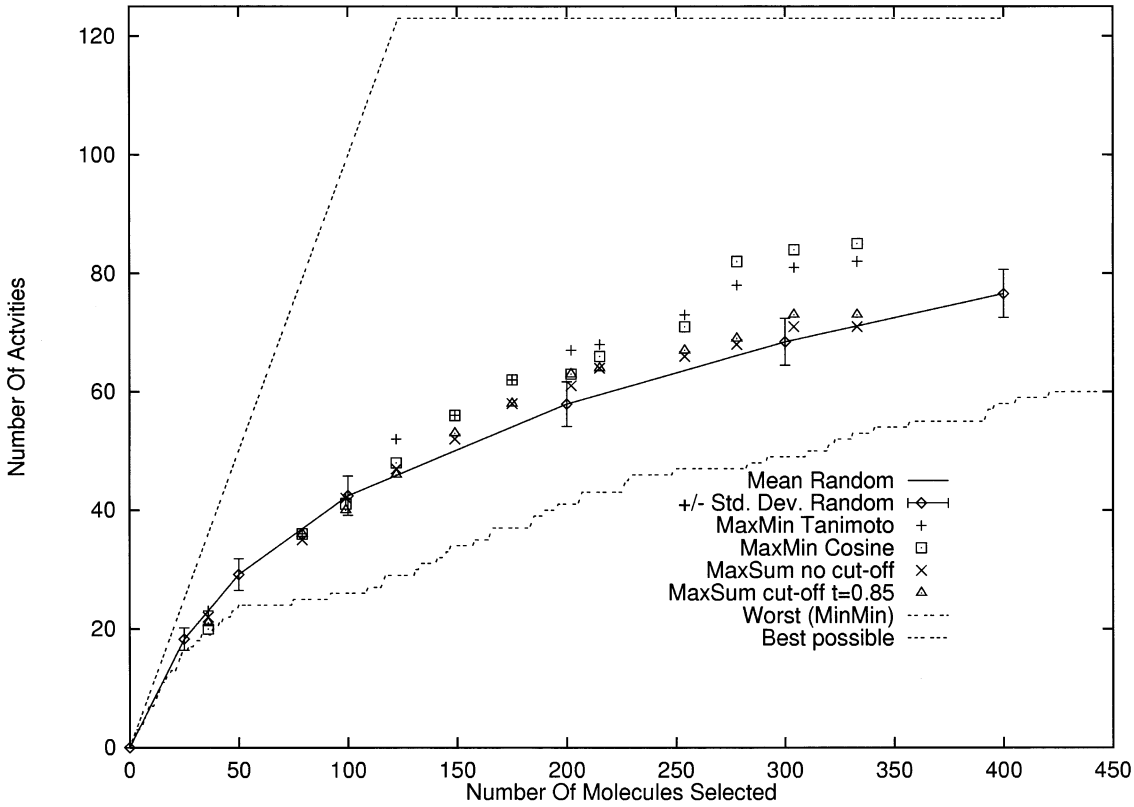

Figure 5. Numbers of activity classes identified using maximum-dissimilarity algorithms with the bit-string data.

algorithms select about the same number of (or fewer) activity classes as by random selection; for larger subsets, MaxMin, whether using the Tanimoto or the cosine coefficient, consistently selects the largest number of activity classes. MaxSum picks only a few more activities than by random selection in this range of subset sizes, and the inclusion of the similarity cutoff, which ensures that no pair of compounds has a Tanimoto dissimilarity less than 0.15, brings little improvement. Broadly comparable results are found when the property data are used, as shown in Figure 6, with the exceptions that there are more subrandom results here than when the bit-string representations are used and that the general superiority of MaxMin is more evident. There are no runs with a cutoff of 0.85 here, since no appropriate cutoff values have been reported in the literature for use with property data.

A few MaxMin runs were carried out using the Euclidean distance as the dissimilarity measure. Noticeably fewer activity classes were selected than when either of the two association coefficients were used, a finding that is in line with previous studies of similarity coefficients for property prediction.[27] Tests with several other association coefficients yielded no results that were significantly superior to those obtained with the Tanimoto or cosine coefficients, and the remaining experiments hence used just these two coefficients.

A previous comparison of MaxMin and MaxSum (and of two other maximum-dissimilarity algorithms)[11] concluded that it was not possible to say that one was superior to the other. However, that comparison involved performance measures that were based on bit-string data, rather than on actual bioactivity data as here, where the results demonstrate the general superior performance of MaxMin as compared with MaxSum.

As noted previously, the size of the subset cannot be specified in a sphere-exclusion algorithm. However, the use of a fixed starting molecule and a deterministic selection procedure for the SE-MinSum, SE-MinMax, or SE-MinMin algorithm means that each of these will always lead to a constant subset. Conversely, the random nature of the MDISS selection procedure means that it is not possible to predict the size or the constitution of the subsets resulting from its use *a priori*. Figure 7 shows the effects of variations in *t*, the threshold dissimilarity, on the size of the subset that is produced, and it will be seen that the various algorithms tend to produce subsets of about the same size, with the possible exception of SE-MinMin, which seems to select marginally fewer molecules at a given sphere radius than do the other algorithms. Figure 7 refers to the bit-string data set; similar results are obtained with the property data set.

Figures 8 and 9 are the sphere-exclusion equivalents of Figures 4 and 5, with each dot representing a single run of the MDISS algorithm. All the selections are better than random for subsets containing more than about 125 molecules, but there is no single algorithm here that consistently selected the greatest number of activity classes. However, the best individual sphere-exclusion results came from individual MDISS runs, and Figure 10 hence shows the Tanimoto MaxMin results superimposed on the MDISS results. It will be seen that MaxMin with the bit-string data selects the same number of activities per subset as one would expect on average from MDISS, while MaxMin with the property data selects about the same number of activities per subset as the best of the MDISS runs (to minimise the number of figures, we have plotted the

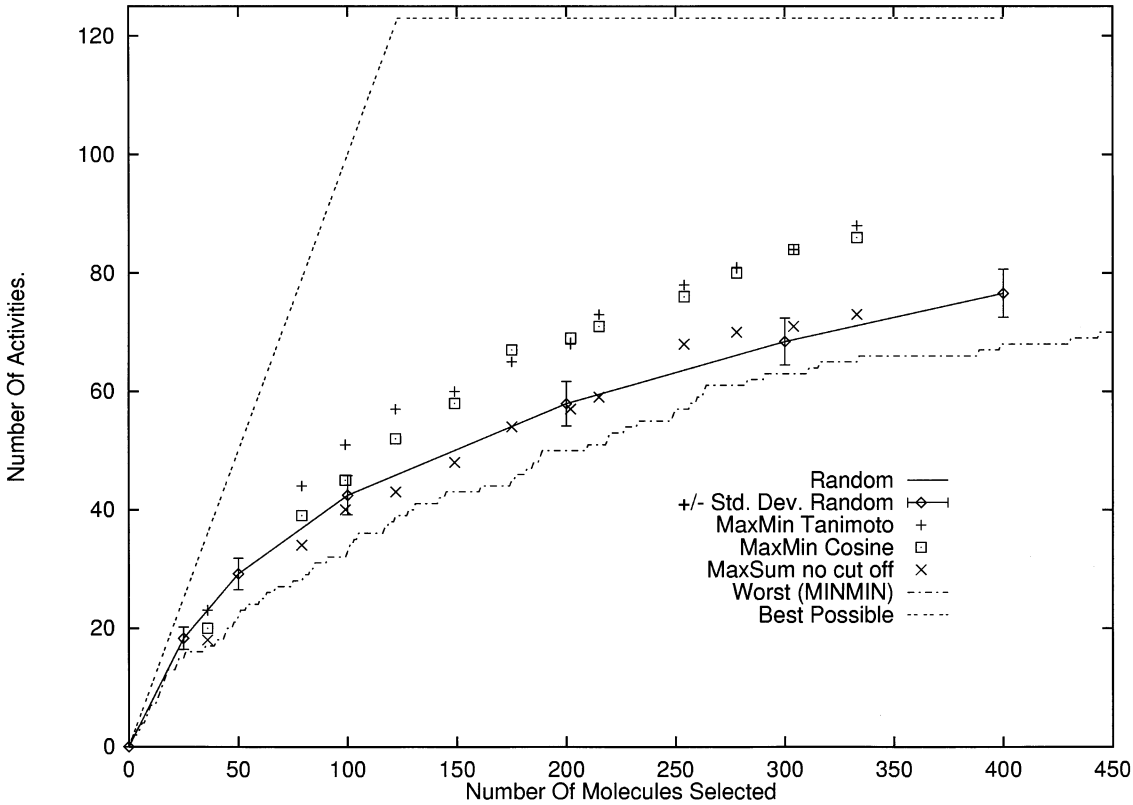

Figure 6. Numbers of activity classes identified using maximum-dissimilarity algorithms with the property data.

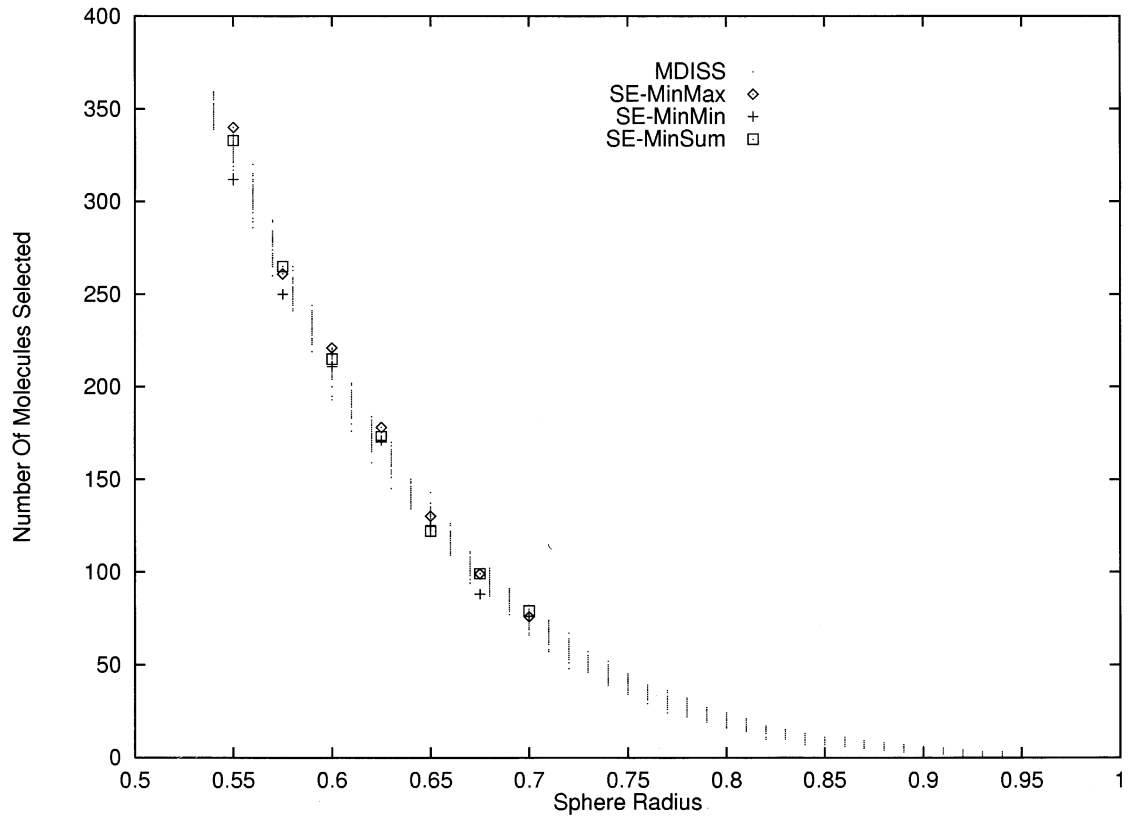

Figure 7. Effect of variations in the sphere radius on the size of the subsets produced by the various sphere-exclusion algorithms with the bit-string data.

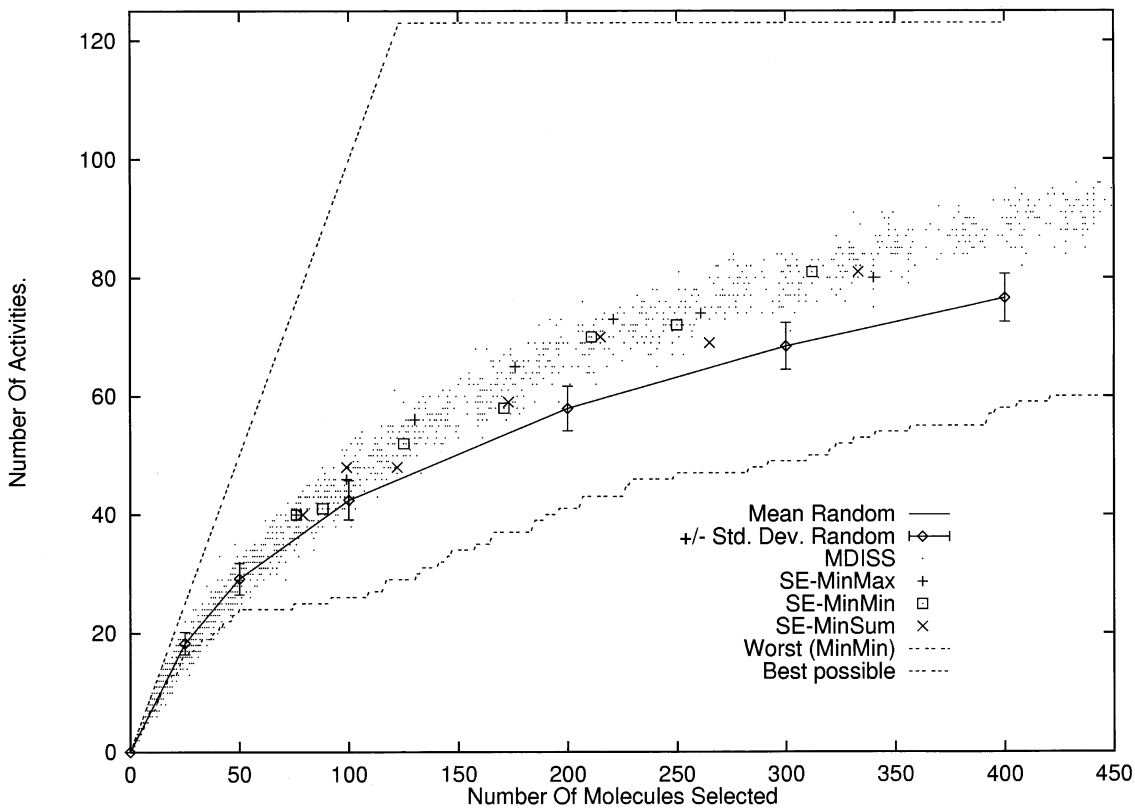

Figure 8. Number of activity classes identified using sphere-exclusion algorithms with the bit-string data.

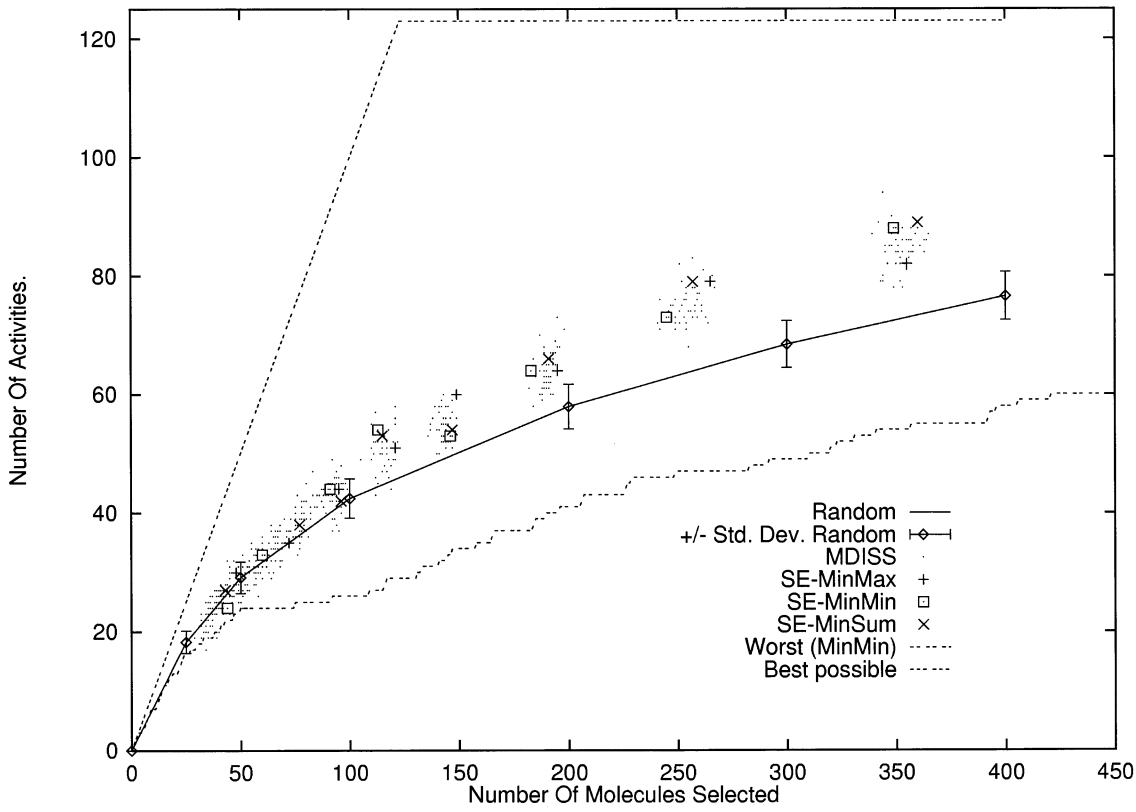

Figure 9. Number of activity classes identified using sphere-exclusion algorithms with the property data.

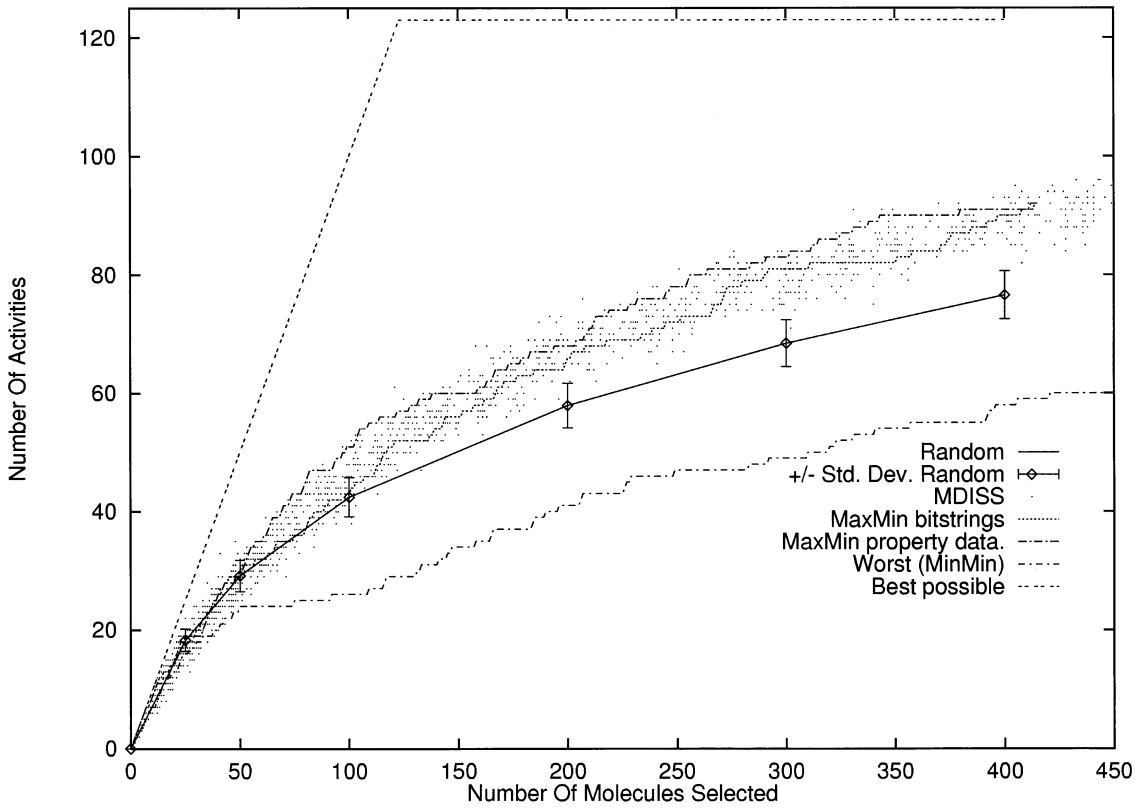

Figure 10. Comparison of the number of activity classes identified using the MaxMin and MDISS algorithms.

MDISS results for both types of data on the same graph, as there was very little difference between them).

Figure 11 focuses upon the nature of the activity classes that are selected, rather than the number, as previously. Figure 11a–c shows the distribution of occurrence frequencies for a subset containing 261 compounds selected from the bit-string data set by MaxSum (without the cutoff), MaxMin, and SE-MinMax. The MaxMin and SE-MinMax distributions are quite similar to each other, and differ from the MaxSum distribution, most notably for those activity classes that have many representatives in the subset. To quantify this, the original data set and each of the selected subsets were regarded as 123-element vectors in which the *i*th element contained the number of compounds belonging to the *i*th activity class. It was then possible to calculate the weighted Tanimoto coefficient between pairs of these vectors. These similarities are shown in Table 1, which demonstrates clearly the variant behaviour of the MaxSum algorithm. Similar conclusions may be drawn from the frequency distributions for the property data set.

To obtain further insights into the workings of the various algorithms, we have randomly generated a data set containing 100 two-dimensional points in [0 . . . 80, 0 . . . 100], and then applied MaxMin, MaxSum, and SE-MinSum to select 10-member subsets from this data set. This data set, and the order in which points are selected, are shown in Figure 12. The first selection by MaxSum is the molecule that is most dissimilar to the rest of the data set, this representing one of the points nearest a "corner" of the data set. It next selects the point that is most dissimilar from the first, i.e., one in the opposite corner, and then continues to select points from the periphery, as shown in Figure 12a. MaxSum thus embodies a bias towards compounds near to the edge of the data set, and will select those nearer to the centre only when all of the outlying regions have been sampled; moreover, it is clear that the omission of a cutoff, as here, can lead to a fair degree of clustering in the subset that is chosen. This behaviour is noticeably different from the selection behaviour of MaxMin and SE-MinSum, as shown in Figure 12b and c, respectively. MaxMin initially selects the same points as MaxSum, but differs once points from the four corners have been selected, with points having maximum minimum dissimilarities relative to those already selected being distributed reasonably evenly across the space. With the exception of MDISS, the sphere-exclusion algorithms are designed to start with the molecule nearest to the centre and then to move outwards only after excluding all molecules from that region of space. While there are differences in the various dissimilarity definitions used for sphere exclusion, they should all provide a reasonable sampling of the entire space when run to completion, i.e., until there are no further molecules that can be selected using the specified dissimilarity radius. If, instead, a sphere-exclusion algorithm is terminated by specifying some fixed number of compounds there is the possibility of a biased sample; for example, an inspection of Figure 12c shows that the selection of just six molecules would result in none being selected from the right-hand portion of the space.

It must be emphasised that the data set used in Figure 12 is very different from the bit-string and property data sets, and we have thus undertaken a further set of experiments with the results shown in Figure 13. Here, a mean vector is calculated for our 2 049-compound data set, and then the complement of the Tanimoto coefficient calculated between this "pseudomolecule" and each of the molecules as they are selected for

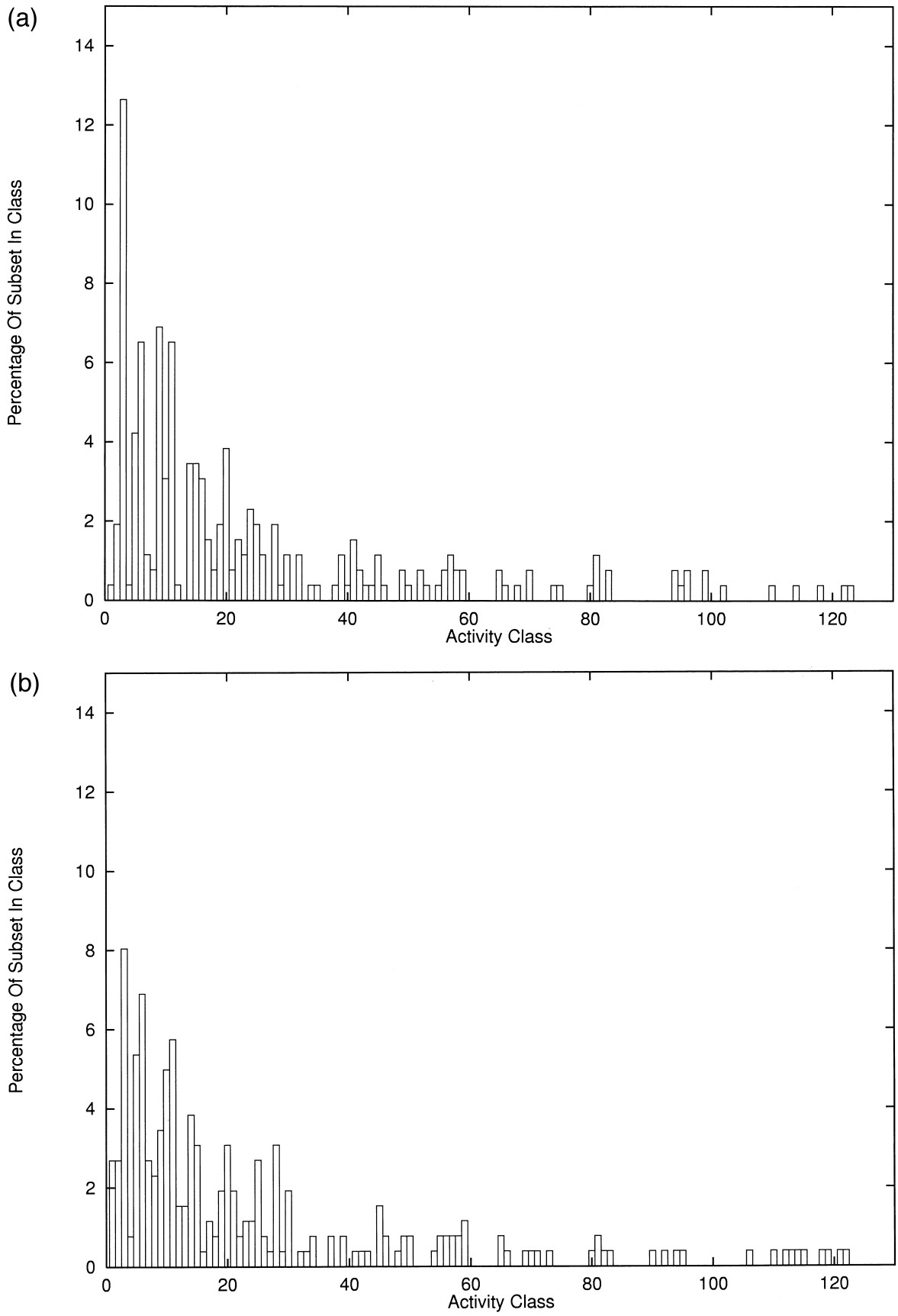

Figure 11. Distribution of bioactivity classes in subsets of 261 compounds selected using (a) MaxSum, (b) MaxMin, and (c) SE-MinMax algorithms (continued).

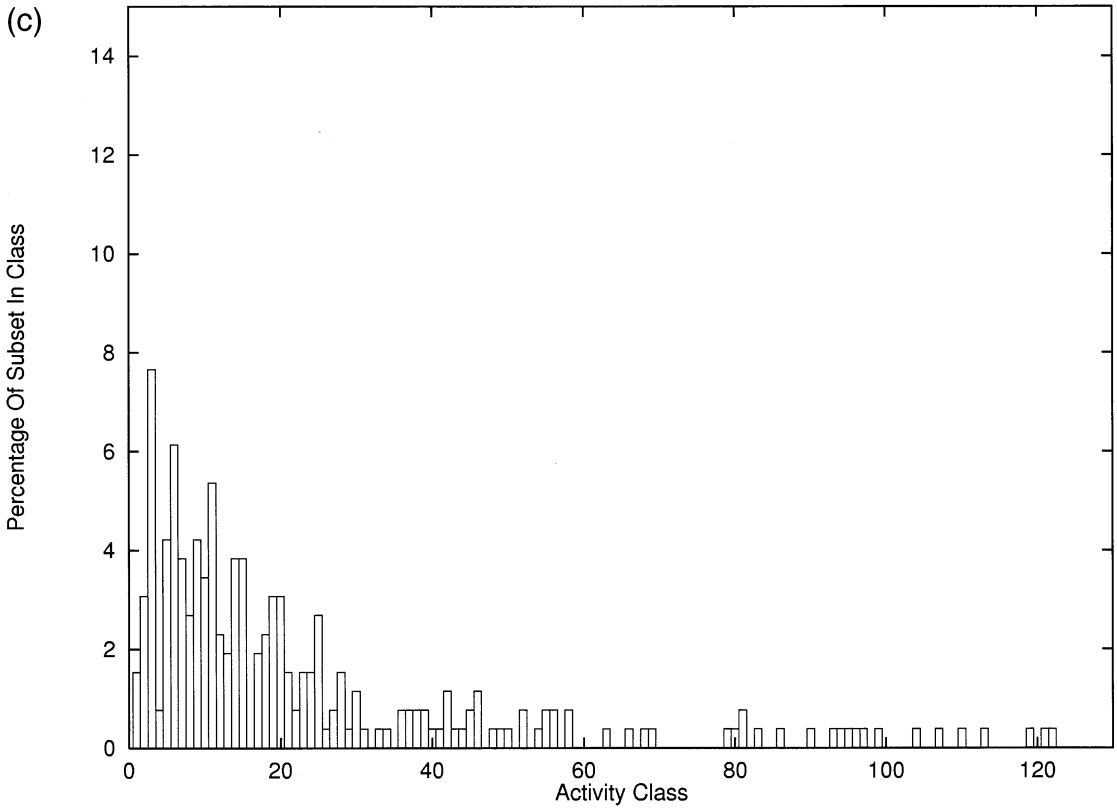

Figure 11. (continued)

inclusion in *Subset*. Figure 13 plots the mean dissimilarity between the pseudomolecule and all of the molecules that have been selected thus far at each iteration of one of the algorithms. Figure 13 shows that the dissimilarity between the central pseudomolecule and the next compound selected by MaxSum decreases as more compounds are selected, thus confirming that the MaxSum algorithm starts by picking compounds far removed from the centre of the data set, and then moves inwards as further molecules are chosen. MaxMin also selects far-distant compounds initially, but then moves closer to the centre (with a consequent lowering of the plotted line when compared with MaxSum), while the SE-MinMax plot demonstrates clearly that this algorithm starts with compounds near to the centre and then moves outwards. These plots are for the bit-string data but similar results are obtained with the property data.

Our results hence suggest that the MaxSum algorithm has an inherent bias towards the selection of outlier compounds, and that the resulting subsets often yield a lower number of different activity classes in our *WDI* data set than do the other selection methods. However, as Taylor notes,[20] such a bias may represent an appropriate selection strategy if the distribution of activity in the data set is such that these outliers have a higher *a priori* probability of activity than do the remaining compounds, or if a premium is placed, as it normally is in a corporate environment, on the identification of activity in novel structural classes.

Further insights into the behaviour of the various algorithms are obtained by calculating the mean nearest neighbour dissimilarity (when averaged over all of the compounds in a subset) to ascertain the extent to which the various algorithms had succeeded in keeping even the least dissimilar molecules in a subset as well separated as possible. In general, MaxMin was found to yield subsets for which the mean nearest neighbour dissimilarity was greater than for the sphere-exclusion algorithm, where the subsets were, in their turn, better separated than for MaxSum. For example, when a 410-compound subset was chosen, the mean dissimilarities were as follows: MaxMin, 0.572; SE-MinSum, 0.553; MaxSum, 0.400. Thus, while MaxSum tends to focus on outlier compounds, it still manages to include molecules that have nearest neighbours that are, on the average, closer than with the other selection algorithms. In this respect, there are similarities with the information-theoretic diversity measure of Lin[31], which has been criticised by Agrafiotis.[32]

Thus far, we have considered the subsets that the various algorithms identify in the database as a whole. We now discuss the results that were obtained in the feedback experiments, in

Table 1. Weighted Tanimoto coefficient values for the distribution of compounds among the 123 activities when comparing selected subsets with each other or with the original complete data set

	MaxMin	MaxSum	SE-MinMax
Data set	0.645	0.486	0.649
MaxMin		0.812	0.918
MaxSum			0.790

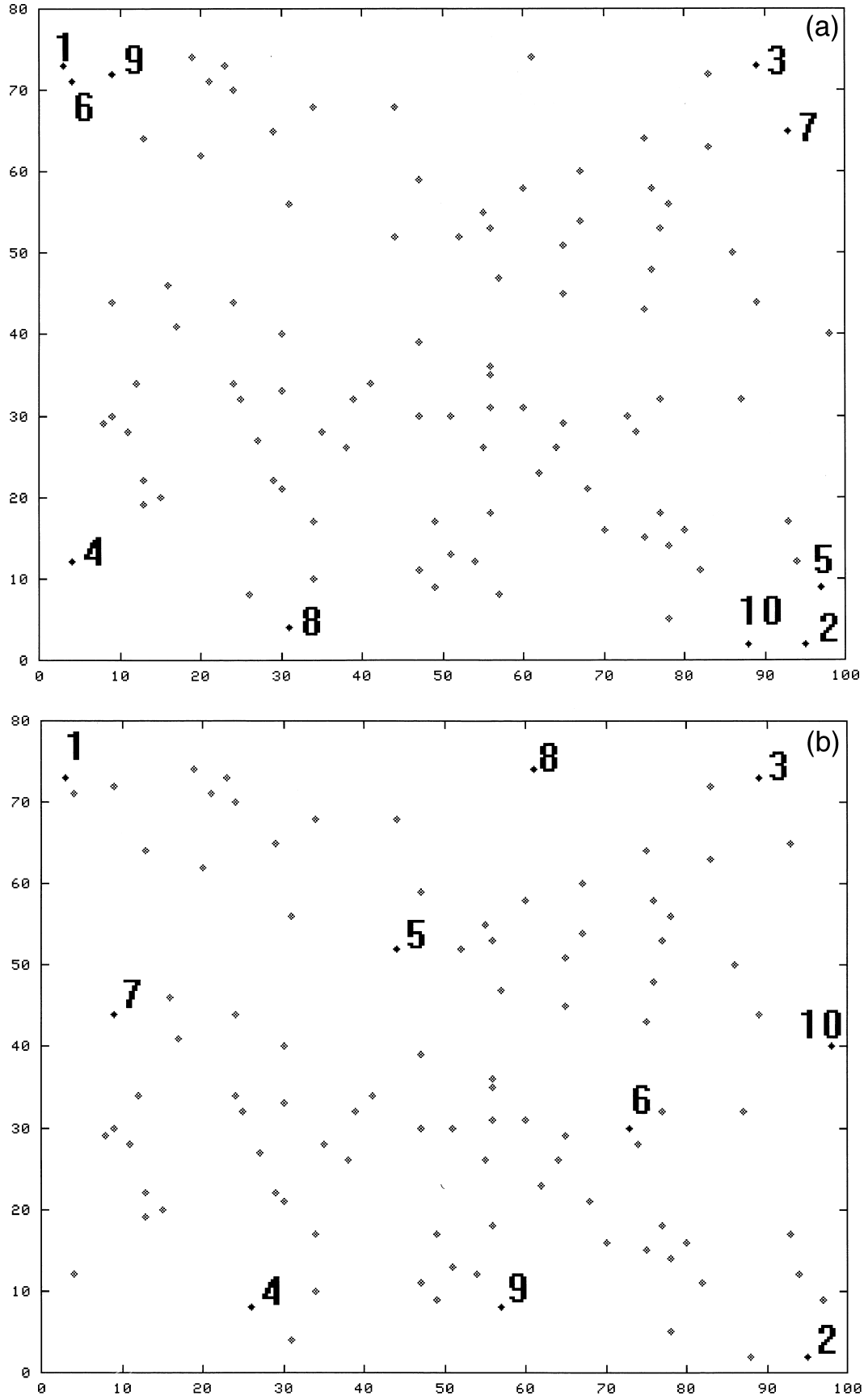

Figure 12. Selection of points from a two-dimensional data set using (a) MaxSum, (b) MaxMin, and (c) SE-MinSum algorithms (continued).

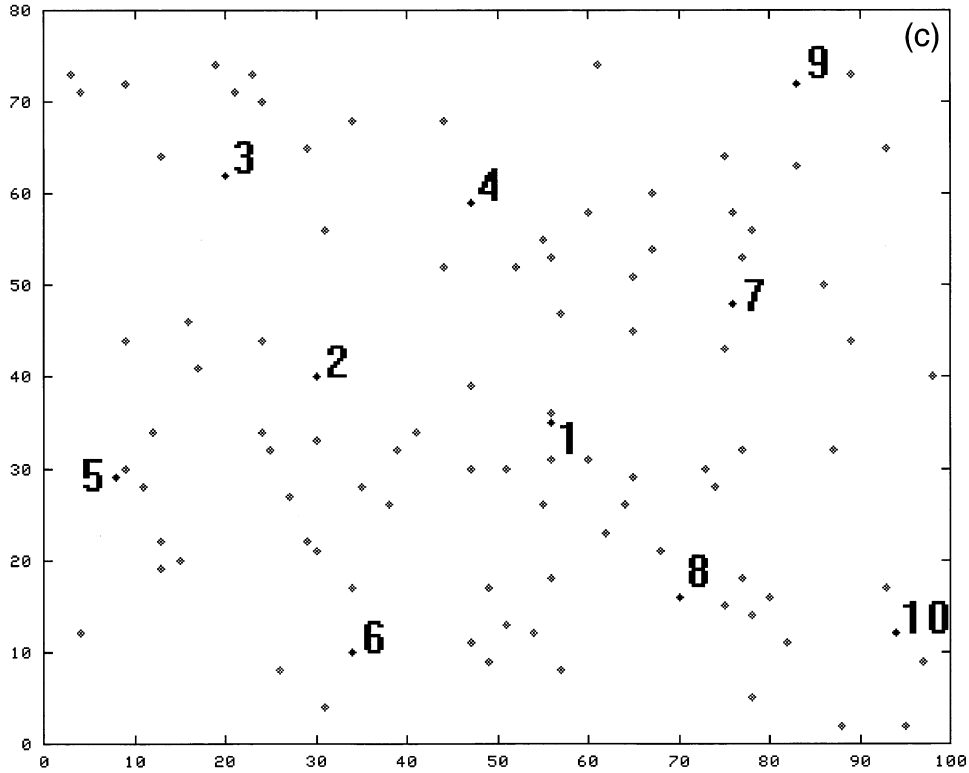

Figure 12. (continued)

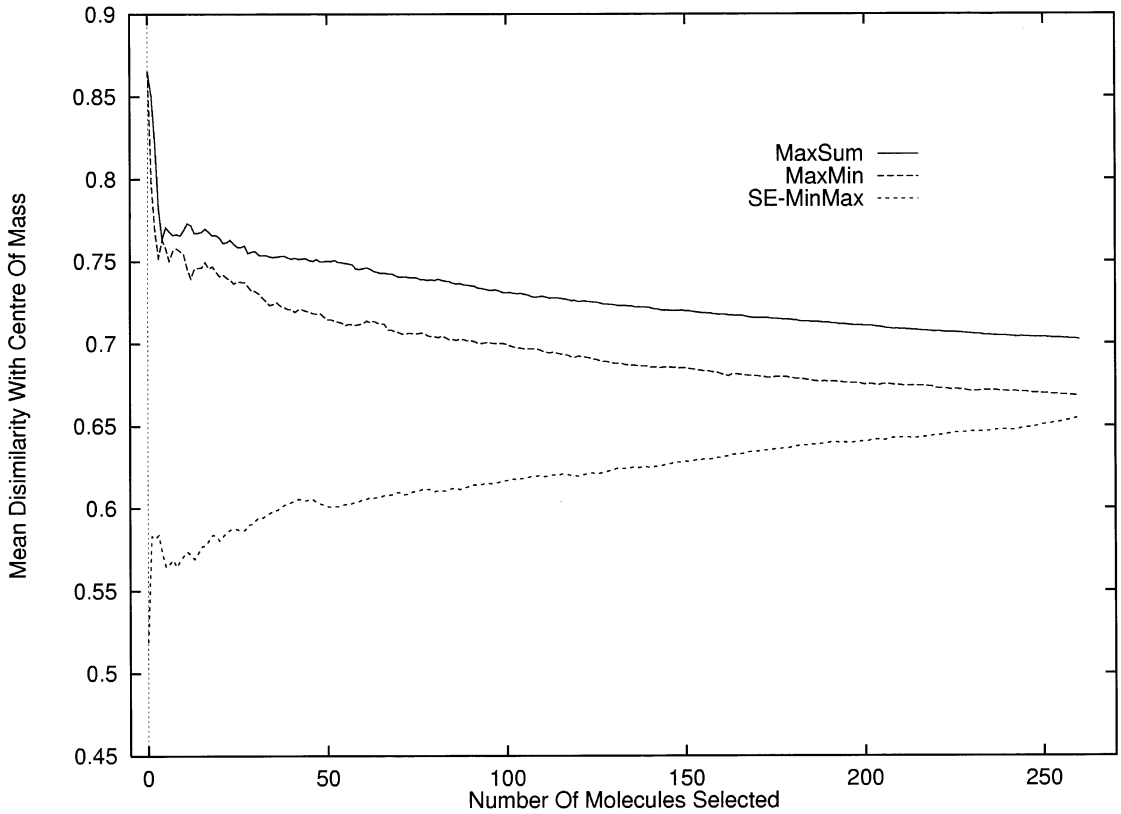

Figure 13. Mean distance from the centre of the data set of the compounds selected using MaxSum, MaxMin, and SE-MinMax algorithms.

Table 2. Feedback experiments[a]

Activity class	MaxSum			MaxMin			SE-MinSum		
A. 265-Compound Subset									
Antibiotics	1	0	0	7	6	0	5	2	0
Analgesics	5	2	0	7	2	0	8	0	0
Cytostatics	33	5	0	22	1	0	22	12	0
Corticosteroids	1	0	0	4	2	0	3	0	0
Antiarterosclerotics	11	1	0	14	1	1	14	0	1
Anthelmintics	17	3	1	18	1	1	13	1	0
Spasmolytics	3	0	0	7	0	0	6	3	2
Antihistamines	2	0	1	6	0	0	6	1	0
B. 410-Compound Subset									
Antibiotics	2	11	0	11	9	0	12	13	0
Analgesics	9	2	0	17	2	0	19	5	0
Cytostatics	43	5	0	33	2	0	38	2	0
Corticosteroids	5	8	0	3	2	0	3	1	0
Antiarterosclerotics	11	1	0	14	1	1	14	0	1
Anthelmintics	17	3	1	18	1	1	13	1	0
Spasmolytics	3	0	0	7	0	0	6	3	2
Antihistamines	2	0	1	6	0	0	6	1	0
C. 529-Compound Subset									
Antibiotics	5	12	0	19	13	0	16	7	0
Analgesics	10	2	0	23	3	0	25	2	0
Cytostatics	47	5	0	42	3	0	44	6	0
Corticosteroids	10	24	0	5	2	0	3	3	0
Antiarterosclerotics	21	2	7	25	2	1	25	2	1
Anthelmintics	27	5	1	28	1	1	28	1	1
Spasmolytics	11	1	1	11	2	0	11	2	0
Antihistamines	9	0	2	16	1	1	16	1	1

[a]The three integers in each element Table 2 represent the number of actives identified in the original subset, and then the numbers of actives and of inactives in the set of feedback compounds.

which the actives identified in a subset were inspected to find how many of their near neighbours were also active. Subsets were generated for the following selection algorithms: MaxSum with a cutoff applied to ensure that no two selected molecules had an intermolecular dissimilarity of less than 0.15 (the feedback results were worse if this cutoff was not applied); MaxMin; and SE-MinSum. The two maximum-dissimilarity algorithms were used to generate subsets containing 265, 410, and 529 molecules, with comparably sized subsets for the SE-MinSum algorithm being obtained with radii of 0.575, 0.520, and 0.480, respectively. The results obtained are shown in Table 2.

Each element in the main body of Table 2 contains three integers: the number of actives identified in the original subset, and then the numbers of actives and of inactives in the set of feedback compounds. An inspection of Table 2 demonstrates that only a very few inactive compounds were retrieved in any of the feedback searches, thus demonstrating the effectiveness of the 0.15 threshold as a filter for eliminating inactive near neighbours of a known active (although there are also, of course, very many actives that are removed by this filter). It is more difficult to make any firm statements as to the relative merits of the three selection algorithms tested, since MaxSum would seem to be less noticeably inferior to the other two algorithms than was the case in the previous sets of experiments. Here, MaxSum successfully identified a large number of actives in these feedback searches, but it also retrieved more inactives than the other two approaches; moreover, it was starting from a lower level, in that there were fewer actives in the original subsets, and it could thus be argued that it had more scope for improvement than did the other two selection methods.

CONCLUSIONS

There is much current interest in the development of methods for selecting structurally diverse subsets of files of chemical structures. In this article, we have compared several different algorithms that have been suggested for dissimilarity-based compound selection and that are all sufficiently rapid in execution for use with large files of compounds. Using a data set drawn from the *World Drugs Index*, our results suggest that algorithmic approaches to compound selection (such as DBCS) are equivalent to random selection when small database subsets are to be identified but are increasingly superior as larger subsets are required. The MaxMin maximum-dissimilarity algorithm is the most effective, of those tested here, in selecting compounds associated with a range of bioactivities; its performance is sometimes exceeded by that of the MDISS algorithm, but the random element in the latter procedure means that its effectiveness varies substantially from one run to the next. As well as being effective in operation, the availability of an $O(nN)$ expected time for the MaxMin algorithm means that it can be implemented on very large data sets without the need for excessive computational resources. We hence conclude that MaxMin would appear to be the best algorithm currently available for nonfocused, dissimilarity-based compound selection.

ACKNOWLEDGMENTS

We thank Pfizer Central Research (UK) for funding, Derwent Information for provision of the *World Drugs Index*, Tripos Inc. for software support, and Val Gillet for helpful discussions on the subject of molecular diversity. This article is a contribution from the Krebs Institute for Biomolecular Research, which has been designated as a centre for biomolecular sciences by the Biotechnology and Biological Sciences Research Council.

REFERENCES

1 Willett, P. Using computational tools to analyse molecular diversity. In: *Combinatorial Chemistry: A Short Course* (S.H. DeWitt and A.W. Czarnik, eds.). American Chemical Society, Washington, 1997, pp 17–48
2 Lajiness, M.S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Design* 1997, **718,** 65–84
3 Johnson, M.A., Lajiness, M.S., and Maggiora, G. Molecular similarity: A basis for designing drug screening programs. In: *QSAR: Quantitative Structure–Activity Relationships in Drug Design* (J.L. Fauchere, ed.). Alan R. Liss, New York, 1989, pp. 167–171
4 Lajiness, M.S., Johnson, M.A., and Maggiora, G. Imple-

menting drug screening programs using molecular similarity methods. In: *QSAR: Quantitative Structure–Activity Relationships in Drug Design* (J.L. Fauchere, ed.). Alan R. Liss, New York, 1989, pp. 173–176
5 Lajiness, M.S. Molecular similarity-based methods for selecting compounds for screening. In: *Computational Chemical Graph Theory* (D.H. Rouvray, ed.). Nova Science, New York, 1990, pp. 299–316
6 Bawden, D. Application of two-dimensional chemical similarity measures to database analysis and querying. In: *Concepts and Applications of Molecular Similarity* (M.A. Johnson and G.M. Maggiora, eds.). John Wiley & Sons, New York, 1990, pp. 65–76
7 Lajiness, M.S. An evaluation of the performance of dissimilarity selection. In: Silipo, C. and *QSAR: Rational Approaches to the Design of Bioactive Compounds* (C. Silipo and M.S. Lajiness, eds.). Elsevier Science, Amsterdam, 1991, pp. 201–204
8 Marengo, E., and Todeschini, R. A new algorithm for optimal, distance-based experimental design. *Chemometrics Intell. Lab. Syst. 1992*, **16,** 37–44
9 Bawden, D. Molecular dissimilarity in chemical information systems. In: *Chemical Structures Vol. 2*: *The International Language of Chemistry* (W.A. Warr, ed.). Springer-Verlag, Heidelberg, 1993, pp. 383–388
10 Holliday, J.D., Ranade, S.S., and Willett, P. A fast algorithm for selecting sets of dissimilar structures from large chemical databases. *Quant. Struct.–Activity Relationships* 1995, **14,** 501–506
11 Holliday, J.D., and Willett, P. Definitions of 'dissimilarity' for dissimilarity-based compound selection. *J. Biomol. Screening* 1996, **1,** 145–151
12 Hudson, B.D., Hyde, R.M., Rahr, E., and Wood, J. Parameter based methods for compound selection from chemical databases. *Quant. Struct.–Activity Relationships* 1996, **15,** 285–289
13 Polinsky, A., Feinstein, R.D., Shi, S., and Kuki, A. LiBrain: Software for automated design of exploratory and targeted combinatorial libraries. In: *Molecular Diversity and Combinatorial Chemistry*: *Libraries and Drug Discovery* (I.M. Chaiken and K.D. Janda, eds.). American Chemical Society, Washington, D.C., 1996, pp. 219–232
14 Gillet, V.J., Willett, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 731–740
15 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., and Weinberger, L.E. Neighbourhood behaviour: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 1996, **39,** 3049–3059
16 Brown, R.D., and Martin, Y.C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 572–584
17 Brown, R.D., and Martin, Y.C. The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 1–9
18 Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* 1997, **40,** 1219–1229
19 Ferguson, A.M., Patterson, D.E., Garr, C.D., and Underiner, T.L. Designing chemical libraries for lead discovery. *J. Biomol. Screening* 1996, **1,** 65–73
20 Taylor, R. Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J. Chem. Inf. Comput. Sci.* 1995, **35,** 59–67
21 Young, S.S., Farmen, M., and Rusinko, A. Random versus rational. Which is better for general compound screening? At http://www.awod.com/netsci/Science/Screening/feature09.html (visited June 4, 1997)
22 Spencer, R.W. Diversity analysis in high throughput screening. *J. Biomol. Screening* 1997, **2,** 69–70
23 Kennard, R.W., and Stone, L.A. Computer aided design of experiments. *Technometrics* 1969, **11,** 137–148
24 Tripos. *DiverseSolutions User's Manual*. St. Louis, Missouri, Tripos, Inc., 1996
25 UNITY is available from Tripos, Inc., at URL http://www.tripos.com/
26 Molconn-Z is available from eduSoft, LC at URL http://www.eslc.vabiotech.com/
27 Willett, P., and Winterman, V. A comparison of some measures of inter-molecular structural similarity. *Quant. Struct.–Activity Relationships* 1986, **5,** 18–25
28 Johnson, M.A., and Maggiora, G.M. (eds.). *Concepts and Applications of Molecular Similarity*. John Wiley & Sons, New York, 1990
29 Dean, P.M. (ed.). *Molecular Similarity in Drug Design*. Chapman & Hall, Glasgow, 1994
30 Downs, G.M., and Willett, P. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* 1996, **7,** 1–66
31 Lin, S.K. Molecular diversity assessment: Logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguisability after rejection of Gibbs paradox of entropy mixing. *Molecules* 1996, **1,** 57–67
32 Agrafiotis, D.K. On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 576–580