ELSEVIER

Binning schemes for partition-based compound selection

Martin J. Bayley and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield, UK

Partition-based approaches to the selection of structurally diverse sets of compounds involve allocating compounds to the individual elements of a multidimensional grid that spans the available chemical space. The space is defined by an appropriate set of chemical properties, with subranges of the values of these properties being used to define the constituent elements, or bins. This article compares several binning schemes in terms of their ability to provide an even distribution of compounds across the available space and to maximise the numbers of active molecules identified in simulated assay experiments. © 1999 by Elsevier Science Inc.

INTRODUCTION

The introduction of combinatorial synthesis and high-throughput screening has led to dramatic changes in the lead discovery programmes undertaken in the pharmaceutical and agrochemical industries. The vastly increased numbers of compounds that are now available for synthesis and testing has spurred the development of a range of computational techniques to maximise the cost-effectiveness of lead discovery.[1–10] Particular attention has been given to the development of methods for the selection of structurally diverse subsets of chemical databases, both real and virtual, using approaches based on clustering, dissimilarity analysis, combinatorial optimisation, and partitioning, the subject of this article.

Partition-based compound selection requires the identification of a set of P characteristics that span the chemical space of interest,[11] these characteristics normally being molecular properties that would be expected to affect binding at a receptor site. The range of values for each such characteristic, I ($1 \leq I \leq P$), is subdivided into a set of B_I subranges, and the combinatorial product of all possible subranges then defines the set of *bins*, or *cells*, that make up the partition. When a subset of a data set is required, each molecule in that data set is assigned to the bin that matches the set of binned characteristics for the chosen molecule; a structurally diverse subset is then obtained by selecting one (or some small number) of the molecules from each of the bins. Partition-based selection can be much faster than cluster, dissimilarity, and optimisation approaches to compound selection, and the availability of a partition enables the explicit identification of those sections of structural space that are underrepresented, or even unrepresented, in a database.[12]

The most obvious factor to be considered in partition-based selection is the set of characteristics that is used to define the chemical space. The first report of such an approach was by Mason et al.,[13] who generated partitions defined by six global molecular properties that had been chosen to encode the hydrophobicity, polarity, hydrogen bond donor and acceptor power, torsional flexibility, and shape of a molecule. However, any sort of global property can be used to generate a partition, such as topological indices,[14] BCUT parameters encoding atomic charges, atomic polarisabilities and atomic hydrogen-bonding abilities,[12] and three-point pharmacophores.[15] In this article, we consider another factor, this being the *binning scheme*, i.e., the algorithm that is used to define the subranges specifying the partition from which compounds are selected. The selection of an appropriate binning scheme is related to the selection of an efficient file structure for database management systems that are designed to handle multidimensional partial-match queries, and we have drawn on this work (specifically on file structures such as the grid file[16] and the *k-d* tree[17]) in the development of the binning schemes considered here.

BINNING SCHEMES

The binning schemes considered here satisfy two simple criteria. First, the maximum and minimum values for each of the characteristics that specify the partition must be set so as to encompass all of the compounds that may need to be processed by the partitioning scheme; second, it seems appropriate that each molecule be assigned to just a single bin, thus requiring that the bin ranges do not overlap at all. These criteria are satisfied by using a simple n-dimensional grid in which each dimension is subdivided by means of $n - 1$ dimensional hyperplanes, e.g., a three-dimensional grid can be subdivided by a set of interlocking two-dimensional planes. There are, however, two ways in which these hyperplanes can be generated. The simpler, *descriptor-independent* partitioning scheme

Address reprint requests to: Peter Willett, University of Sheffield, Krebs Institute for Biomolecular Research, Western Bank, Sheffield S10 2TN, UK. Tel.: +44-114-222-2633; Fax: +44-114-278-0300; E-mail: p.willett@sheffield.ac.uk

Journal of Molecular Graphics and Modelling 17, 10–18, 1999

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00016-9

assumes that the hyperplane used to subdivide the *I*th dimension is completely independent of the hyperplanes that have been used to subdivide the preceding $I - 1$ dimensions. Alternatively, a *descriptor-dependent* partitioning scheme generates a hyperplane to subdivide the *I*th dimension by taking account of the hyperplanes that have been used to subdivide the preceding $I - 1$ dimensions. The compounds that find themselves within a particular bin are determined by the bin ranges (i.e., the ranges of values for each of the *n* characteristics defining a bin) that are adopted. Two simple criteria present themselves for controlling these ranges (and hence the occupancy of each bin): splitting the descriptor space of each dimension into equally *sized* subdivisions or splitting the descriptor space of each dimension into equally *occupied* subdivisions.

It is hence possible to identify four types of binning scheme, depending on whether a hyperplane is, or is not, independent of its predecessors and on the occupancy criterion that is used to define each of the bins. The combination of these two factors specifies four different binning schemes, as illustrated in Figures 1–4 for a simple two-dimensional data set. Thus, in Figure 1 bin boundaries are identified for descriptor 1 such that each of the resulting ranges contains approximately the same number of molecules; the same boundary selection procedure is then involved for descriptor 2, but without taking any account of the previously identified descriptor 1 boundaries. In Figure 2, the range of values for, e.g., descriptor 1, is subdivided into an appropriate number of equisized ranges, with, again, no account being taken of these bin boundaries when descriptor 2 is considered. In Figures 3 and 4, conversely, the bin boundaries for descriptor 2 are determined by the specific boundaries that have been identified previously for each subdivision of descriptor 1. Thus, in Figure 3, the descriptor 2 ranges resulting from application of the equifrequency criterion to the first descriptor 1 subdivision (shown shaded in dark grey) are quite different from the descriptor 2 ranges that become associated with the second descriptor 1 subdivision (shown shaded in lighter grey). Figure 4 illustrates the use of the equisized (rather than the equifrequent) criterion with such dependent hyperplanes.

PERFORMANCE MEASURES

A comparison of the effectiveness of the four binning schemes described in the previous section requires quantitative performance measures, and we have used a total of four measures for this purpose. The first two measures seek to quantify the occupancy, i.e., the distribution of compounds throughout the bins that are specified by a scheme, and the second two measures seek to quantify the performance of the partition when used for lead discovery purposes.

The first occupancy-based performance measure is the well-known statistic, χ^2, as suggested previously by Pearlman and Smith.[12] Assume that a data set containing *N* compounds has been allocated to a partition containing *B* bins, and that the *I*th bin contains *O*(*I*) compounds. Then *E(I)*, the expected number of compounds in each bin, *I*, is given by *N*/*B* and the χ^2 statistic is given by Eq. (1):

$$\chi^2 = \sum_{I=1}^{B} \frac{[E(I) - O(I)]^2}{E(I)} \tag{1}$$

The second occupancy measure is the Pratt measure of class concentration.[18] Define *q* by means of Eq. (2):

$$q = \sum_{I=1}^{B} \frac{R(I) \times O(I)}{N} \tag{2}$$

where, as before, *O*(*I*) is the number of compounds in the *I*th bin and where *R*(*I*) is the rank of the *I*th bin when the bins are ranked in decreasing order of their *O*(*I*) values. Then the class concentration, *C*, is given by Eq. (3):

$$C = \frac{B + 1 - 2q}{B - 1} \tag{3}$$

In both cases, the smaller the value of the measure, the more even the distribution of compounds across the partition.

The two occupancy-based performance measures require just a set of compounds that has been partitioned by some binning scheme: the second group of performance measures requires in addition that the compounds have some associated biological activity data. The first such measure takes as its basis the assumption that the best possible partition would be one in which the binning was such as to ensure that all of the active molecules in a data set were completely isolated from the inactive molecules. Let an *active bin* be one that contains at least one active molecule, then the optimal partition would be

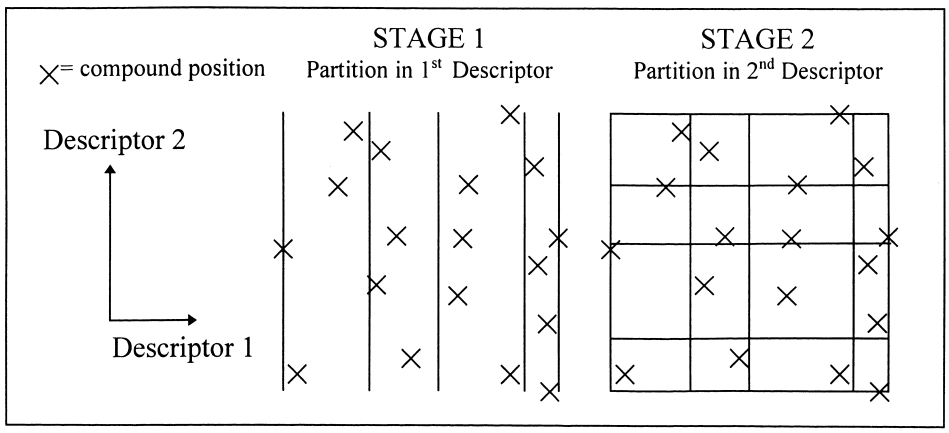

Figure 1. A 2D representation of the boundaries of an equifrequent, independent binning scheme.

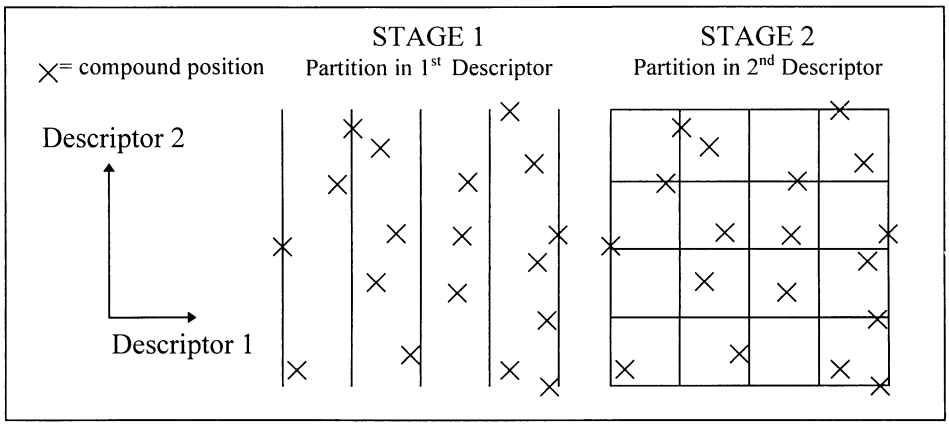

Figure 2. A 2D representation of the boundaries of an equisized, independent binning scheme.

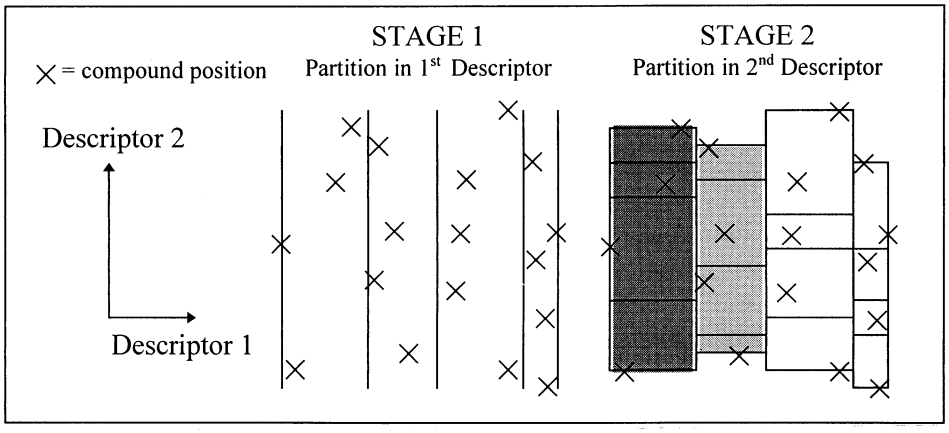

Figure 3. A 2D representation of the boundaries of an equifrequent, dependent binning scheme.

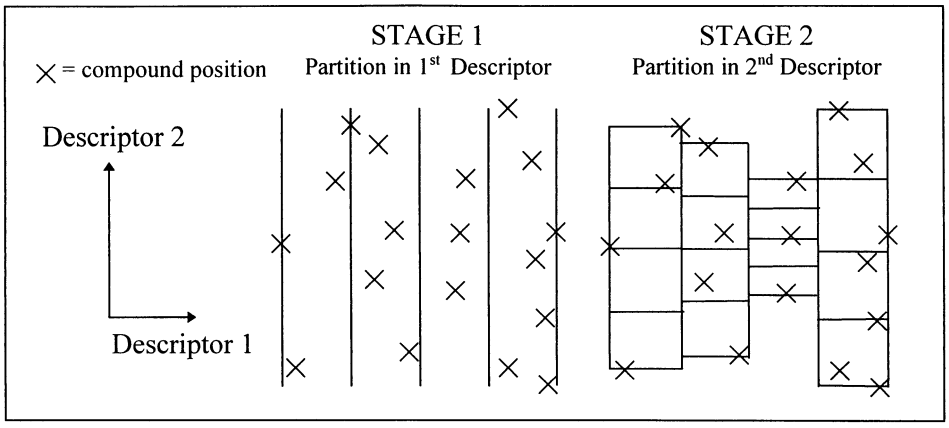

Figure 4. A 2D representation of the boundaries of an equisized, dependent binning scheme.

one in which the active bins did not contain any inactive molecules. A measure of the deviation, *D*, from this ideal situation, was obtained by summing the number of inactive compounds found within the active bins and then dividing this sum by the total number of active molecules. *D* is thus the number of inactive molecules binned with each active molecule, i.e., it describes the occurrence of false positives in the active bins, and the higher its value the less effective the partitioning. A similar measure, but based on the occurrence of actives, rather than inactives, in active bins has been described by Brown and Martin.[19]

The final measure estimated the performance of a particular

partition when used in an operational lead discovery programme. It is assumed that a partition has been generated for a data set and that the molecules in the data set have been allocated to their appropriate bins. One compound is selected at random from each bin in turn and checked to see if it is active (in an actual discovery programme, this would correspond to that compound being assayed): if this proves to be the case then all of the other molecules in that bin are also checked for activity. We refer to this two-stage procedure as *bin expansion.* Assume that a total of T compounds is checked in the two stages, i.e., the initial set of B compounds from the B bins and then the other members of the active bins. Then the performance measure used was the ratio of the number of actives, P, identified via the partition to the number of actives, E, that would be expected to be obtained if T compounds were selected at random. (Note that if some bins in a partition do not contain any molecules then the procedure will result in less than B compounds being selected in the first stage. In such cases, the nonempty bins were systematically resampled until the required number of compounds had been selected for checking.) Assume that an N-compound data set contains a total of n actives. Then E is given by Eq. (4):

$$E = \frac{T \times n}{N} \tag{4}$$

and the performance ratio, R, is given by Eq. (5):

$$R = \frac{\Sigma P}{\Sigma E} \tag{5}$$

where the sum is over all of the distinct activity classes associated with each of the three data sets discussed in the next section (1 class in the case of the NCI and TRR data sets, and 15 in the case of the WDI data set).

EXPERIMENTAL DETAILS AND RESULTS

Three data sets were used in our experiments. The first of these (WDI) contained 5 145 uncharged structures from the World Drug Index.[20] Each of the molecules in this database has one or more associated broad-class activity indicants, and our tests used the molecules with 1 or more of 15 commonly occurring indicants. The second data set (NCI) contained 3 508 (92 active and 3 416 inactive) molecules drawn from the publicly available National Cancer Institute AIDS database of compounds that have been tested in an antiviral screen for HIV inhibition.[21] The third data set (TRR) was provided by Tripos Receptor Research Limited[22] and consisted of 2 143 compounds for which a proprietary binding assay had been performed: a cutoff was applied to the measured percentage inhibitions to yield a total of 60 active compounds.

The compounds in each of these three data sets were characterised in two different ways. First, the molecules in each data set were input to the MOLCONN-Z program[23] for the calculation of a total of 178 topological indices. These indices were then input to a principal components analysis (PCA) and the molecules were represented by the first four principal components resulting from this analysis. These components described 81, 77, and 78% of the variance in the data for the WDI, NCI, and TRR data sets, respectively. The order in which the components were identified specified the order in which the dependent partitions were created. Thus, the first component formed the first dimension, which was divided into the required number of subranges. Each of these subranges then provided the basis for the subdivision of the second dimension, i.e., second principal component, and so on until the bins had been fully specified. Second, the molecules in each data set were characterised by some of the properties used in the biological activity profiles described recently by Gillet et al.[24]; here, each molecule was described by the number of hydrogen bond donors (or the number of rotatable bonds in the case of the TRR data set), the number of hydrogen bond acceptors, the number of aromatic rings, and the kappa-2 shape index. In what follows we shall use the notation *x–y* to denote a particular combination of data set and representation, where *x* is WDI, NCI, or TRR and where *y* is 1 (for the MOLCONN-derived descriptions) or

Table 2. *D* occupancy measure for each binning scheme

Data set	Equifrequent dependent	Equifrequent independent	Equisized dependent	Equisized independent
WDI–1	9.15	12.33	28.35	36.84
WDI–2	9.74	14.98	36.32	42.56
NCI–1	5.25	7.54	18.50	27.32
NCI–2	5.94	8.41	32.57	36.85
TRR–1	6.01	8.94	21.67	29.16
TRR–2	6.54	13.20	18.98	27.94

Table 1. *C* and χ^2 occupancy measures for each binning scheme

	Equifrequent dependent		Equifrequent independent		Equisized dependent		Equisized independent	
Data set	*C*	χ^2	*C*	χ^2	*C*	χ^2	*C*	χ^2
WDI–1	0.035	35.2	0.508	5384.3	0.811	41947.1	0.921	94875.2
WDI–2	0.035	35.4	0.529	6752.8	0.875	63467.1	0.931	94448.4
NCI–1	0.017	6.9	0.428	2191.6	0.816	37917.8	0.951	134890.6
NCI–2	0.017	7.4	0.501	4090.3	0.954	88017.9	0.997	497076.8
TRR–1	0.059	41.8	0.380	1021.2	0.814	15601.7	0.930	30992.8
TRR–2	0.059	42.3	0.635	3985.9	0.787	9369.8	0.928	29089.4

Table 3. Bin expansion results for the three data sets

Data set	Binning scheme: Equifrequent or equisized	Binning scheme: Dependent or independent	Select	Found	*R*
WDI–1	Equifrequent	Dependent	99.42	21.29	7.84
		Independent	150.82	32.87	7.22
	Equisized	Dependent	551.56	42.11	3.41
		Independent	1 031.48	58.37	2.70
WDI–2	Equifrequent	Dependent	97.93	21.42	7.77
		Independent	158.22	25.09	6.16
	Equisized	Dependent	916.24	50.78	2.32
		Independent	1 200.37	61.39	2.17
NCI–1	Equifrequent	Dependent	85.29	16.70	7.47
		Independent	124.13	18.92	5.81
	Equisized	Dependent	566.84	36.99	2.49
		Independent	1 401.19	51.30	1.40
NCI–2	Equifrequent	Dependent	86.67	20.43	8.99
		Independent	234.27	33.59	5.47
	Equisized	Dependent	919.42	26.66	1.11
		Independent	2 749.54	71.86	1.00
TRR–1	Equifrequent	Dependent	52.51	2.83	1.96
		Independent	77.96	2.43	1.13
	Equisized	Dependent	383.45	11.46	1.09
		Independent	639.05	17.07	0.97
TRR–2	Equifrequent	Dependent	51.76	1.42	1.00
		Independent	155.18	6.40	1.50
	Equisized	Dependent	219.11	7.47	1.24
		Independent	533.92	16.01	1.09

2 (for the biological activity profile descriptions); thus, NCI–1, for example, refers to the MOLCONN-derived representations of the AIDS compounds.

Each of the binning schemes was applied to each combination of data set and representation, using each of the performance measures described in the previous section. The results of these experiments are detailed in Tables 1–3, which refer to partitions containing 256 bins, with each of the four dimensions being divided into four subranges.

Tables 1 and 2 show the results of the occupancy experiments, which describe the distribution of compounds across the available bins. It will have been realised from the preceding discussion that the equifrequent, dependent partitions most closely model the distribution of compounds in the input data set, and it is thus hardly surprising to find that this binning scheme results in the lowest Pratt and χ^2 values: in the ideal case, i.e., if the number of compounds was an exact multiple of 256 and if none of them had equal sets of values, both of these performance measures would be zero for this particular scheme. The other three schemes all give very much larger values, the equisized ones particularly so. The figures hence demonstrate that an equifrequent scheme will give a more even distribution of compounds than will an equisized scheme, and that a dependent scheme will give a more even distribution than an independent one. These results are hardly surprising, but entirely comparable conclusions can be drawn from the *D* values listed in Table 2, which take activity information into account. The values listed here for WDI–1 and WDI–2 are mean values obtained by averaging over all 15 of the WDI activity classes; however, the same pattern is observed if the results for each of the individual activity classes are considered, and we have thus merely presented the average results in the interests of brevity.

A similar averaging process has been used in the presentation of the bin expansion experiments in Table 3. As noted previously, these experiments involved simulating the selection of 256 compounds, 1 from each bin in a partition, and then the checking of all compounds in the active bins. The column headed "Select" here contains the numbers of compounds in those active bins, the column headed "Found" is the number of additional active compounds identified once the selected compounds have been checked for activity, and *R* is the performance ratio defined previously, with larger values of *R* denoting better performance. An inspection of this final column shows the same behaviour as was observed in the occupancy experiments, with the sole exception of the TRR–2 dependent results (although none of the *R* values here are large).

One of the advantages of partition-based selection is that the bins provide an obvious way of identifying further candidates for testing once an active molecule has been obtained, namely the other molecules in that bin. Similar comments apply to cluster-based selection,[1] where the clusters can play a role similar to that of the bins here, but dissimilarity-based selection[25] requires a time-consuming similarity search to be carried

Table 4. Bin expansion results for the three data sets, using random selection

Data set	Binning scheme: Equifrequent or equisized	Binning scheme: Dependent or independent	Select	Found	*R*
WDI–1	Equifrequent	Dependent	87.00	16.97	7.39
		Independent	160.03	29.60	7.00
	Equisized	Dependent	537.87	39.97	3.33
		Independent	1 018.33	57.05	2.67
WDI–2	Equifrequent	Dependent	87.28	16.74	7.16
		Independent	148.13	22.09	5.89
	Equisized	Dependent	903.19	50.00	2.32
		Independent	1 189.56	60.92	2.17
NCI–1	Equifrequent	Dependent	73.77	13.80	7.13
		Independent	113.05	16.54	5.58
	Equisized	Dependent	548.45	33.15	2.31
		Independent	1 382.30	48.62	1.34
NCI–2	Equifrequent	Dependent	73.06	15.95	8.32
		Independent	224.89	31.95	5.42
	Equisized	Dependent	906.84	25.03	1.05
		Independent	2 746.92	71.11	0.99
TRR–1	Equifrequent	Dependent	46.33	2.48	1.94
		Independent	71.98	2.27	1.14
	Equisized	Dependent	377.60	11.29	1.09
		Independent	631.67	16.94	0.97
TRR–2	Equifrequent	Dependent	45.49	1.23	0.98
		Independent	149.60	6.14	1.49
	Equisized	Dependent	211.27	6.99	1.20
		Independent	524.81	15.68	1.08

Table 5. Comparison of partition-based[a] and random-based bin expansion, using MOLCONN-Z descriptions

Data set	Systematic		Random		*Z*
Corticosteroids	10.50,	0.89	10.13,	1.19	25.47
Estrogens	6.31,	1.82	5.86,	1.96	17.11
Antihistamines (H1)	4.32,	1.47	4.24,	1.57	3.54
Sympatholytics (β)	5.07,	1.96	4.85,	1.98	8.06
Progestogens	10.82,	2.10	10.68,	2.38	4.29
Dopamine antagonists	4.39,	1.89	4.37,	1.94	0.89
Antiserotonins	4.50,	2.20	4.44,	2.26	1.69
Sympatholytics (α)	7.33,	4.73	6.92,	4.89	6.09
Tranquilizers	3.61,	2.99	3.52,	2.97	2.02
Angiotensin antagonists	3.86,	2.19	3.84,	2.23	0.69
Phospodiesterase inhibitors	3.81,	2.61	3.85,	2.66	−1.09
Androgens	19.32,	7.02	18.68,	7.62	6.20
Antihistamines (H2)	7.90,	5.50	7.82,	5.55	0.95
Serotoninergics	12.63,	12.50	12.20,	12.69	2.43
Gabaminergics	12.29,	9.80	12.46,	10.03	−1.23
NCI	7.37,	2.43	7.16,	2.68	5.96
TRR	2.01,	1.08	1.96,	1.19	3.11

[a] Using the equifrequent–dependent scheme.

Table 6. Comparison of partition-based[a] and random-based bin expansion, using activity profile descriptions

Data set	Systematic		Random		*Z*
Corticosteroids	9.20,	1.24	8.40,	1.53	40.67
Estrogens	8.65,	2.98	7.98,	3.21	15.24
Antihistamines (H1)	6.51,	1.96	6.33,	2.04	6.25
Sympatholytics (β)	16.00,	3.29	15.24,	3.90	14.76
Progestogens	3.29,	1.30	3.30,	1.32	−0.98
Dopamine antagonists	3.30,	2.10	3.29,	2.20	0.14
Antiserotonins	3.93,	2.62	3.84,	2.72	2.41
Sympatholytics (α)	3.03,	2.02	2.98,	2.07	1.67
Tranquilizers	3.83,	2.55	3.75,	2.56	2.38
Angiotensin antagonists	4.24,	2.32	4.23,	2.44	0.50
Phospodiesterase inhibitors	14.05,	6.93	13.49,	7.17	5.58
Androgens	4.61,	5.27	4.61,	5.33	0.00
Antihistamines (H2)	17.72,	22.91	14.97,	22.82	8.50
Serotoninergics	13.54,	10.22	12.39,	10.54	7.82
Gabaminergics	3.83,	2.55	3.75,	2.56	2.38
NCI	8.96,	3.20	8.28,	3.54	14.27
TRR	0.99,	0.74	1.00,	0.83	−1.18

[a] Using the equifrequent–dependent scheme.

Table 7. Calculated *Z* scores for various numbers of bins *B*, using the equifrequent–dependent scheme for WDI–1 data sets

	B				
Data set	16	81	256	625	1296
Corticosteroids	5.47	17.14	23.22	41.69	36.87
Estrogens	4.08	7.48	16.82	38.37	34.44
Antihistamines (H1)	2.13	2.19	5.37	8.91	7.69
Sympatholytics (β)	2.48	2.84	8.23	8.94	11.79
Progestogens	3.78	2.44	4.76	14.66	13.80
Dopamine antagonists	1.48	0.06	2.54	1.35	*−2.12*
Antiserotonins	−1.95	2.66	0.59	1.99	*−10.28*
Sympatholytics (α)	−0.11	2.91	6.90	3.38	15.21
Tranquilizers	0.96	0.63	2.23	3.48	−0.87
Angiotensin antagonists	3.37	0.28	−1.06	0.60	1.89
Phosphodiesterase inhibitors	0.84	0.33	1.04	8.03	21.67
Androgens	4.58	5.61	4.20	12.37	14.58
Antihistamines (H2)	0.95	0.73	1.18	−0.70	−0.94
Serotoninergics	*−3.60*	2.92	2.78	3.99	19.24
Gabaminergics	*−7.13*	−0.35	0.53	6.79	*−3.28*
NCI	1.64	1.95	8.16	8.88	12.70
TRR	−1.26	0.79	3.02	*−2.54*	−0.37

out if expansion is required. This is normally effected by setting some similarity threshold (e.g., a value of 0.85 for the Tanimoto coefficient in a fingerprint-based similarity search[19]) and then retrieving all molecules having a similarity exceeding this threshold with any of the active molecules identified in the first stage. It is thus of interest to note that the use of bin expansion as described here was found to give *R* values that were not substantially inferior to those obtained by using the actives identified in the first stage as the target structures for a series of similarity searches. Thus, the bins provide a means for expansion that is slightly less effective than a similarity search but that is far more efficient, the two expected time complexities being $O(1)$ and $O(N)$, respectively.

There has been some controversy as to the effectiveness of systematic procedures for the selection of compounds.[26–29] It is thus of interest to compare the results obtained here with those obtained from random selection. The *R* values demonstrate a clear increase in the number of active compounds identified when compared with the numbers of active compounds that would be expected from random selection from the data set as a whole, with the possible exception of the TRR data sets. A more stringent test of the effectiveness of the partitions is to compare the *R* values obtained from expansion of the active bins when the initial set of 256 compounds is obtained by systematic sampling of the partition, with the corresponding values obtained from expansion of the active bins when the initial set of 256 compounds is selected randomly. The results obtained from such a procedure are presented in Table 4. This shows, first, a similar ordering of the binning schemes as is revealed by Table 3; and, second, that the *R* values here are generally smaller than those in Table 3, i.e., that the systematic selection procedure gives slightly better results than does the random selection procedure. It will be seen that the results in Tables 3 and 4 for the TRR data sets are generally smaller than for the data sets from the other two, evaluated sources, and it would thus be of interest to repeat the experiments reported here with additional such data.

Table 8. *C* and χ^2 occupancy measures for sample-based ranges in the equifrequent–dependent scheme, using WDI-2 data set

Percent	*C*	χ^2
100	0.035	35.4
50	0.174	492.2
20	0.323	1 860.9

The latter finding is studied in more detail in Tables 5 and 6, using the individual activity classes for the WDI data set and using the equifrequent-dependent partitions, as this scheme had given the best results in the previous tests. For each such partition, a representative compound was selected at random from each of the 256 bins and then the expansion ratio, *R*, calculated. This was repeated 10 000 times and the mean and standard deviation calculated for each of the distinct activities, as listed in the column marked "Systematic" in Table 5 (which involved the MOLCONN-Z descriptions) and Table 6 (which involved the biological activity profile descriptions). The results in the columns in these tables marked "Random" were obtained as for Table 4, i.e., by selecting 256 compounds at random to initiate the expansion (rather than one from each of the 256 bins as in partition-based selection). The procedure was again repeated 10 000 times and the means and standard deviations for *R* calculated. A *Z* test for the comparison of two means was then carried out to determine whether there was any statistically significant difference between the sets of values resulting from the systematic and random procedures. Given

two distributions of sizes n_1 and n_2, and with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, Z is calculated as shown in Eq. (6):

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}} \quad (6)$$

The critical value for Z in a two-tailed test with $p \leq 0.05$ is 1.96, and there are thus significant differences in performance for all but the underlined entries in the column marked Z in Tables 5 and 6.

Results similar to those presented above are obtained if different-sized partitions are employed. Thus, Table 7 lists the Z values obtained using the equifrequent-dependent scheme and the MOLCONN-Z descriptors with grids containing 16 (2^4), 81 (3^4), 625 (5^4), and 1 296 (6^4) cells. In addition to the many underlined, i.e., not significant, values, there are also some significant negative values for Z, i.e., where the random results were, on average, better than the partition-based results; these are both underlined and italicised. There is some tendency for the effectiveness of the partition-based results to increase as more discriminating grids are employed, but there are also notable exceptions, as with the antiserotonins.

A limitation of equifrequency-based binning is that the partitions are data set dependent. The final set of experiments hence involved using only some of the molecules in a data set, specifically the WDI–2 data set, for the generation of the equifrequent-dependent ranges. The complete data set was then allocated to the bins in the normal way, and the occupancy measures calculated. These results are shown in Table 8, which considers random samples containing either 20 or 50% of the entire data set. While poor when compared with the results for the full data set processed by this binning scheme, as shown in Table 1, these sample values are still greater than those obtained for the full data set using any of the other three binning schemes.

CONCLUSIONS

In this article we have discussed binning schemes for the specification of the parameter–value ranges associated with each of the bins in a system for partition-based compound selection. Experiments with three different data sets and two different types of molecular description demonstrate clearly the effectiveness of binning schemes that equalise the numbers of molecules in each bin of a partition, when compared with schemes that equalise the value ranges associated with each of the dimensions of the partition. Our results also provide some support for the use of systematic, as against random, selection procedures.

ACKNOWLEDGEMENTS

We thank the Biotechnology and Biological Sciences Research Council for funding, Derwent Information and Tripos Receptor Research for the provision of data sets, Tripos Inc., for software support, and Val Gillet and David Turner for helpful comments on the work. 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 Shemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., Moreland, D.W., and Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided Mol. Design* 1995, **9**, 407–416

2 Hudson, B.D., Hyde, R.M., Rahr, E., and Wood, J. Parameter based methods for compound selection from chemical databases. *Quant. Struct.-Activity Relat.* 1996, **15**, 285–289

3 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

4 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

5 Willett, P. Using computational tools to analyze molecular diversity. In: *A Practical Guide to Combinatorial Chemistry* (DeWitt, S.H., and Czarnik, A.W., eds.). American Chemical Society, Washington, D.C., 1997, pp 17–48

6 Brown, R.D. Descriptors for diversity analysis. *Perspect. Drug Discov. Design* 1997, **7/8**, 31–49

7 Higgs, R.E., Bemis, K.G., Watson, I.A., and Wikel, J.H. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 861–870

8 Nilakantan, R., Bauman, N., and Haraki, K.S. Database diversity assessment: New ideas, concepts and tools. *J. Comput.-Aided Mol. Design* 1997, **11**, 447–452

9 Clark, R.D. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 1181–188

10 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

11 Mason, J.S., and Pickett, S.D. Partition-based selection. *Perspect. Drug Discov. Design* 1997, **7/8**, 85–114

12 Pearlman, R.S., and Smith, K.M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Design* 1998, **9/10/11**, 339–353

13 Mason, J.S., McLay, I.M., and Lewis, R.A. Applications of computer-aided drug design techniques to lead generation. In: *New Perspectives in Drug Design* (Dean, P.M., Jolles, G., and Newton, C.G., eds.). Academic Press, London, 1994, pp 225–253

14 Cummins, D.J., Andrews, C.W., Bentley, J.A., and Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* 1996, **36**, 750–763

15 Pickett, S.D., Mason, J.S., and McLay, I.M. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* 1996, **36**, 1214–1223

16 Nievergelt, J., Hinterberger, H., and Sevcik, K.C. The grid file: An adaptable, symmetric multikey file structure. *ACM Trans. Database Syst.* 1984, **9**, 38–71

17 Bentley, J.L. Multidimensional binary search trees in database applications. *IEEE Trans. Soft. Eng.* 1979, **SE-5**, 333–340

18 Carpenter, M.P. Similarity of Pratt's measure of class concentration to the Gini index. *J. Am. Soc. Inf. Sci.* 1979, **30**, 108–110
19 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**, 1996, 572–584
20 Information regarding the World Drug Index is available at http://www.derwent.com/
21 Information regarding the National Cancer Institute's AIDS database is available at http://epnws1.ncifcrf.gov:2345/dis3d/aids_screen/aidspub.html
22 Information regarding Tripos Receptor Research Limited is available at http://wavespace.waverider.co.uk/~receptor/
23 Information regarding MOLCONN-Z is available at http://www.eslc.vabiotech.com/
24 Gillet, V.J., Willett, P., and Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 165–179
25 Snarey, M., Terret, N.K., Willett, P., and Wilton, D.J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Model.* 1997, **15**, 372–385
26 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
27 Young, S.S., Farmen, M., and Rusinko, A. Random versus rational. Which is better for general compound screening? At http://www.awod.com/netsci/Issues/Aug96/feature3.html
28 Spencer, R.W. Diversity analysis in high throughput screening. *J. Biomol. Screening* 1997, **2**, 69–70
29 Wikel, J.H., and Higgs, R.E. Applications of molecular diversity analysis in high throughput screening. *J. Biomol. Screening* 1997, **2**, 65–67