Cell-Integral-Diversity Criterion: A Proposal for Minimizing Cluster Artifact in Cell-Based Selections

Obdulia Rabal, Rosalia Pascual, José I. Borrell, and Jordi Teixidó*

Grup d'Enginyeria Molecular, Institut Químic de Sarrià, Universitat Ramon Llull, Via Augusta 390, E-08017 Barcelona, Spain

Received October 10, 2006

Cell-based methods and the diversity integral criterion (a distance-based technique) are commonly used approaches for assessing the diversity of collections of compounds in terms of space coverage. The main deficiency with cell-based methods is the arbitrariness of cell boundaries which leads to edge effects or cluster artifacts, i.e., situations in which similar molecules separated by a cell boundary yield a higher diversity score than molecules falling within the same cell but which are less similar to each other. We describe a straightforward diversity metric based on quantifying the distance to the center of the bins resulting from partitioning the descriptor space which aims at bypassing these artifacts. The mentioned criteria are compared for the diversity assessment of a set of selections carried out on three combinatorial libraries of different cardinalities. For each method, the influence of its parameters (reference partition and number of points) on their efficacy is examined. Furthermore, the proposed diversity metric is also applied to designing diverse libraries for three test cases. We show that full arrays selected by minimizing the sum of distances to the center of the cells are formed by compounds spaced further apart than selections obtained by maximizing the degree of cell occupancy.

INTRODUCTION

The concept of molecular diversity is widely employed in the design of combinatorial libraries aiming to achieve the greatest coverage of the accessible property space. In this diversity-oriented strategy, a limited sample of molecules is selected to maximize the number of chemical and structural features, implicitly assuming, according to the similarity principle,[1] that this process explores the range of the biological activities of the species contained in the library. In database-comparison tasks diversity is used as well to identify unexplored regions, known as "diversity voids", and complement corporate libraries with new members holding differential features.

In recent years, an extensive number of diverse compound-selection methods have been developed, for which several classification schemes have been suggested by different authors.[2−4] Following the classification proposed by Willett[2] and Perez,[3] these methods can be grouped into four broad categories: distance-based methods, also called dissimilarity methods, partition-based or cell-based methods, cluster-based methods, and experimental design methods. Distance-based methods measure diversity as a function of the pairwise intermolecular dissimilarities. Distance is defined by several metrics[5] (Euclidean, Tanimoto, Cosine, Dice, ...), which are implemented in many different functions depending on the dissimilarity concept to maximize (MaxMin,[6,7] Power Sum,[7] Product,[7] Minimum spanning tree,[8,9]...). Cell-based methods divide the descriptor space into hypercubic cells or bins and assign each molecule to the cell that matches the set of binned properties of that molecule.[10,11] Clustering techniques group molecules that show a high degree of both intracluster similarity and intercluster dissimilarity.[12] In these last two categories, diversity is defined in terms of the occupancy of their corresponding partitions (cells or clusters). Several objective functions have been introduced to quantify occupancy which are based on space coverage (number of occupied cells), population coverage (number of compounds in occupied cells), distribution of compounds in cells (χ^2 test), and entropy of the system.[9,11] Finally, the most commonly used experimental design strategy is D-Optimal design.[13]

Another important issue in combinatorial library design is the combinatorial nature of the selected subset. Full array selections, also referred to as sublibraries or arrays, contain products resulting from all combinations of a specific subset of reagents, being the selection performed in the product space. The combinatorial restriction imposed on the selected reagents compels a stochastic optimization. Evolutionary algorithms[14] and simulated annealing[6,15] are by far the most widely used techniques. Conversely, the selection is called cherry picking or sparse array if it just maximizes a desired diversity metric.

Considering the great number of selection methodologies, several criteria have been proposed to judge their effectiveness by assessing the representativeness of the selected subset with respect to the parent library.[6,11,14−18] That is, the extent to which selection spreads in the range of properties offered by the entire library. In that sense, cell-based methods, besides being a selection strategy, provide a simple and efficient mechanism for comparison purposes, their computational cost scaling to the number of compounds in the set $O(N)$. The encoding of absolute position in space makes them portable within different databases characterized by the same set of descriptors and useful for detecting diversity voids.

* Corresponding author phone: +34-932-672-000; fax: +34-932-056-266; e-mail: j.teixido@iqs.url.es.

10.1021/ci600433c CCC: $37.00
Published on Web 09/08/2007

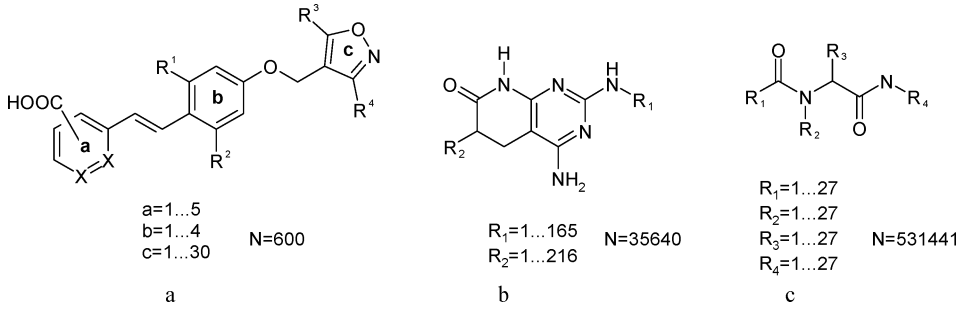

Figure 1. (a) Library I containing 600 (5 × 4 × 30) compounds for FXR partial agonist activity. (b) Two-component 165 × 216 combinatorial library of substituted pyrido[2,3-*d*]pyrimidines (Library II). (c) Ugi virtual library generated from 27 × 27 × 27 × 27 building blocks (Library III).

Additionally, new compounds can be easily added in the analysis of the set. Nevertheless, they are restricted to low-dimensional spaces and are susceptible from suffering from edge effects or cluster artifacts due to the arbitrariness of cell boundaries.

The diversity integral criterion[19] was specially designed to evaluate diversity in library acquisition programs. It is based on comparing the sum of the distances between *p* random points in the common property space of two libraries and the closest molecule to those points from each of these libraries. The library with the lower value samples better the property space and is therefore considered more diverse than the other. The reproducibility of the measurements is restricted to the random character of the chosen points. Furthermore, the use of an inadequate reduced set of points could potentially bias the results.

In the present work, we present a new criterion that we have named cell-integral-diversity criterion which constitutes a compromise between the diversity integral criterion and cell-based methods. As it is the case with the diversity integral function, it is based on the measurement of distances between points and compounds in the descriptor space, but in this case, the points are not placed randomly but in the center of the bins obtained by a partitioning technique. The motivation for developing the present approach stems from the need to avoid the known cluster artifacts of cell-based methods and the disadvantages associated with the random character of the points in the diversity integral criterion, namely reproducibility and the requirement of using a great number of points.

Finally, the three aforementioned criteria (cell-based, diversity integral, and cell-integral-diversity) are also examined and compared for the diversity assessment of a set of selections carried out on three combinatorial libraries. Moreover, we show that the edge effects are diminished when selecting combinatorial subsets using the cell-integral-diversity metric instead of the commonly used cell fraction metric.

METHODS

Data Sets. The first case study (Library I) is a three-component combinatorial oxazole library for partial FXR agonist activity. It consists of 600 compounds resulting from the combination of 5 vinyl-substituted aryl and heteroaryl carboxylates, 4 halophenols, and 30 4-hydroxymethylisoxazoles (Figure 1a). This small library has been previously taken as a reference to analyze selection methodologies for combinatorial library design.[20] It is worth remarking that it has been totally tested against FXR. Thirty-six oxazoles are described to be actives (those with activity at a concentration of 1 μM equal to 50 μM CDCA), which represents 6% of the total library.[21] This has enabled a descriptors' validation,[20] observing a satisfactory neighborhood behavior. For its characterization, a total of 100 descriptors were used: physicochemical, spatial, topological indices, and information content indices. Dimensionality was reduced with Principal Component Analysis (PCA), resulting in seven principal components (PCs) that account for 90% of the variance.

Our second virtual library, named Library II, is a medium size library of substituted pyrido[2,3-*d*]pyrimidines (Figure 1b). This is a two-component library, which is easily accessible by using the Victory multicomponent reaction from R2 substituted acrylates and R1 substituted guanidines.[22] The reagents selection was based on reagent availability. So after substructure search and filtering by both molecular weight and synthetic compatibility the final reagent lists included 216 suitable methacrylic structures and 165 guanidines. This resulted in a 35 640-membered virtual library. After enumeration with Cerius2,[19] geometry optimization was carried out in MOE[23] using MMFF94[24] with a convergence gradient of 0.01 kcal/mol. Characterization was computed with a standard 2D/3D set of 84 descriptors in MOE, including physicochemical descriptors, topological indices based on information theory topological indices, and spatial descriptors.[25] Finally, PCA was applied to reduce dimensionality of the data, and the first nine PCs containing 90% of the variance present in the original descriptors were used to define the chemical space.

The third library (Library III) was based on the Ugi reaction (Figure 1c). For demonstration purposes, 27 acids (R1), 27 amines (R2), 27 aldehydes (R3), and 27 isonitriles (R4) were selected at random from the Maybridge Database. The resulting combinatorial virtual library of 531 441 products was built with Cerius2, optimized, and characterized following the same procedure as described for Library II. The 84 descriptors were reduced to seven principal components, capturing 82% of the total variance of the data set.

Selection Sizes. The degree of agreement between the three diversity criteria was evaluated from two points of view: comparisons of (i) collections of different cardinality within a method and (ii) subsets of the same size selected by different methods. The first approach was applied to Library I and Library II, whereas the second one was tested on Library III.

Thus, for Library I and Library II a total of 21 full array selections were made over a range of selection sizes centered

Table 1. Configuration (Number of Reagents at Each Position) of Full Arrays Selected at Each Selection Size for Each Library

Library I				Library II		
selection size	R1	R2	R3	selection size	R1	R2
15	5	1	3	160	10	16
16	2	2	4	165	11	15
17	1	1	17	170	17	10
18	2	3	3	175	25	7
19	1	1	19	180	12	15
20	2	2	5	185	37	5
21	3	1	7	190	19	10
22	2	1	11	200	10	20
23	1	1	23	205	5	41
24	2	2	6	210	14	15
25	5	1	5	215	43	5
26	2	1	13	220	11	20
27	3	3	3	225	15	15
28	2	2	7	230	23	10
29	1	1	29	235	5	47
30	3	2	5	240	20	12
32	2	2	8	245	35	7
33	3	1	11	250	10	25
34	2	1	17	255	15	17
35	5	1	7	260	20	13
36	3	2	6	265	53	5

around $\sqrt{N}$, with N being the total number of compounds of each library: $\sqrt{600} \approx 25$ for Library I and $\sqrt{35\,640} \approx 189$ for Library II. In Table 1, we show the number of fragments to select for each position depending on the selection size. These configurations (exact number of reagents at each position) were randomly chosen, as it was not the aim of this work to determine the "absolute" best combinatorial configuration at a concrete selection size but a methodological comparison. This, for example, means that for a 36-element selection we randomly selected to use a $3 \times 2 \times 6$ configuration rather than $4 \times 3 \times 3$ or $2 \times 9 \times 2$ configurations.

For library III, a selection size of 1296 compounds was fixed. Subset selections were carried out in the form of 1296-membered sparse arrays as well as $6 \times 6 \times 6 \times 6$ full arrays.

Selections Methods. The following diverse selection methods were explored using the Pralins program (Program for Rational Analysis of Libraries in silico).[20]

(1) Distance-based methods: Among the distance-based diversity selection methods, MaxMin (eq 1) and MaxMin averaged (eq 2) are probably the most commonly used approaches. The former maximizes the minimum intermolecular dissimilarity (D_{ij}), whereas the corresponding objective function for the latter is the mean nearest neighbor distance. In this work, the MaxMin averaged function using Euclidean metric to measure distances was chosen. Subset selections were carried out with a genetic algorithm.

$$\text{MaxMin: } \max\{\min\{D_{ij}^{\,2}\}\} \tag{1}$$

$$\text{MaxMin\_Averaged:}\max\{\sum_{i=1}^{n} \min_{i\neq j;j\subset n} d_{i,j}\} \tag{2}$$

(2) Cell-based methods: The chemical space is partitioned according to the Optimum Binning scheme.[9,20] This algorithm divides the property ranges to create a number of occupied cells or hypercubes less than or equal to the requested number of compounds and with sides as similar as possible. It operates in an iterative fashion, uniformly partitioning the variable (descriptor or PC) with the largest cell edge until the desired number of occupied cells is attained.

(3) Clustering techniques: The K-means relocations algorithm[12] was chosen. This is a nonhierarchical clustering method that starts with an arbitrary set of n compounds acting as cluster centroids and assigns the compounds to the closest centroid. Once all the compounds have been classified into the n groups, the positions of the centroids are recalculated. This process is iteratively repeated until the centroids no longer change.

In these two last methods, the objective function to optimize is the fraction of space covered (cell fraction or cell counts), i.e., the number of cells/clusters occupied by the subset divided by the total number of cells/groups. When designing sparse arrays, a random compound from each cluster or cell is picked.

For the stochastic optimizations (full array design and MaxMin averaged method), we used a genetic algorithm newly implemented in Pralins.[20] The encoding of each potential solution in a chromosome resembles the scheme implemented in the program GALOPED.[26] The individuals are selected, according to their fitness, by the proportional linear method. In order to facilitate convergence, we apply elitism, so a user-defined number of best members are unconditionally passed forward into the next generation. The process is terminated after a preset maximum number of generations or if the best individual score does not improve after a specified number of steps. For all libraries and selection sizes, GA parameters were as follows: population size of 30 individuals, 20 000 iterations to convergence (2000 idle iterations), one-point crossover with a rate of 0.8, a mutation rate of 0.6, and an elitism rate of 3 (10% of population). These parameters were all empirically adjusted after successive trials.

For Library I and Library II, the 21 full array selections (Table 1) were optimized under the MaxMin average function. For Library III, a total of seven subsets were collected coming from the combination of the three methods (MaxMin averaged, cell-based and K-means) with the two designs (cherry picking and $6 \times 6 \times 6 \times 6$ full arrays) plus a random run. In all cases, the Euclidean distance was defined to measure the pairwise dissimilarities.

Diversity Evaluation Criteria. Once the selections were made, we measured their diversity by the three mentioned diversity indices (cell-based, diversity integral, and the proposed cell-integral-diversity). In all three cases, the chemical space was defined by the same set of PCs used for the selection process. All selections and evaluations were done with Pralins[20] and additional Perl scripts to automate the tasks.

Cell-Based Method. The Optimum Binning algorithm described in the previous section was employed to bin the accessible property space. The coverage of each subset was determined as the fraction of occupied cells. Thus, higher values indicate better coverage and greater diversity, resulting in a better representativeness of the available chemical space. Hereafter, we will refer to this metric as SPC. As stated above, a major drawback of the partition methods is the arbitrariness of cell boundaries. To demonstrate this effect and to study the impact on the performance of the chosen partition scheme, six different optimal partitions of the total

Table 2. Six Partition Schemes Employed To Assess the Diversity of the Arrays Selected from the Three Libraries

	Library I		Library II		Library III	
partitioning scheme	no. total cells	no occupied cells	no. total cells	no occupied cells	no. total cells	no occupied cells
P1	24	22	75	74	400	371
P2	40	40	90	76	500	459
P3	80	66	150	147	600	594
P4	120	112	220	213	700	646
P5	150	129	260	254	800	794
P6	200	200	500	442	900	833

space in terms of the total number of bins were used to assess the space coverage of each array. In Table 2, we list the number of total and occupied bins of the six binning schemes applied to each of the three libraries.

Diversity Integral Criterion. Traditionally, this method has been applied to directly compare two subset libraries from a virtual library by dispersing sampling points (probes) over the chemical space of the two subsets. In this study, to cross-compare the diversity of more than two subsets in just a single calculation, we took as reference the total chemical space covered by the parent library. Thus, for a given selection of size *n*, the diversity (D.I) is defined as

$$\mathrm{D.I} = \frac{\sum_{i=1}^{p} \min D_{ij} - \sum_{i=1}^{p} \min D_{ik}}{p} \qquad (3)$$

where p corresponds to the number of sampling points, and D_{ij} and D_{ik} are the Euclidean distances between one random point p and the closest compound from the n-compound subset and the N-membered virtual library, respectively. The sum (integral) is normalized by the number of random points. Therefore, now the lower the value, the more diverse the array is. The number of points is commonly adjusted to ensure a well distributed set, i.e., a uniform sampling. Thus, we initially set the generation of 1000 random points for Library I and 10 000 probes for Library II and Library III. Besides, to test the influence of this parameter on D.I, we ran a series of six calculations handling the same reduced numbers of points that were going to be used in the cell-integral-diversity measurements described in the next section (see Table 2). Each calculation was repeated 10 times.

Cell-Integral-Diversity Criterion. There are plenty of references pointing to the edge effect problems associated with binning and clustering methods.[3,9,20,27] In Figure 2, we exemplify a hypothetical situation of a cluster artifact associated with the cell-based criterion that lead us to propose this index. A virtual library represented by small circles is shown in a two-dimensional descriptor space, wherein two 3-compound subsets A and B are selected. Although a visual inspection shows subset B (*doughnuts*) to be less dispersed than subset A (*squares*), when dividing the chemical space in three cells, subset B and subset A have cell fraction values of 2/3 and 1/3, respectively. Therefore, subset B would be improperly qualified as more diverse. However, using the suggested cell-integral-diversity criterion, the sum of minimum distances from each point located in the center of each bin (*triangles*) to a compound in subset A (a1+a2+a3) is lower than the corresponding value for subset B (b1+b2+b3). Thus, subset A would be correctly considered to sample better the property space than subset B. This already mentioned problem can also be experienced in cluster-based selections. As Patterson and co-workers[27] previously pointed out, there is no guarantee that molecules within a cluster are closer together than molecules in different clusters.

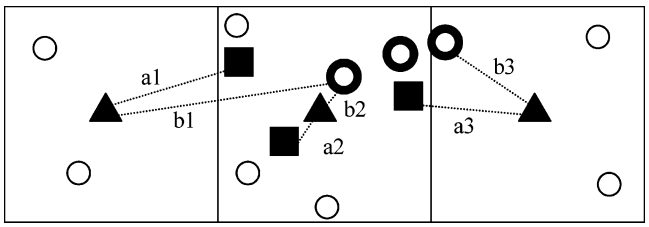

Figure 2. Intended case to represent a cluster artifact. Subset A (*squares*), subset B (*doughnuts*), and points placed into the center of each bin (*triangles*).

Thus, the cell-integral-diversity criterion, like the diversity integral criterion, is based on measuring distances from the probes to the compounds (eq 3), where in this case the probes are positioned in the center of each occupied hypercube generated by the Optimum Binning algorithm. The shorter the distance to the center, the more spread will be the selected subset. As before for D.I, lower values indicate greater coverage. Hereafter, this criterion will be termed as C.I.D. Again, the six partitions listed in Table 2 were taken as frames to locate the points for each subset of the three virtual libraries.

Although the diversity integral criterion and the suggested strategy are very similar in essence, one should keep in mind that the actual property space sampled by these two criteria is not the same. In a multidimensional space defined by *i* properties (x1...x*i*), the random points of the D.I method are positioned in a hypercube defined by coordinates [$x1_{min}$, $x1_{max}$]...[xi_{min}, xi_{max}]. Thus there are sample points in no chemically accessible regions of space in contrast to the cell-integral-diversity method that takes only into account the hypercubic regions around the chemically occupied space. Therefore compounds at the edges of the descriptor space have a greater importance within the D.I method than within the C.I.D index, a factor that might enhance the so-called *edge design*.[28] This behavior is illustrated with the example in Figure 3 for the evaluation of two different 6-membered subsets A and B (*solid dots* in parts a and b, respectively, of Figure 3) selected from a collection of molecules (*unfilled circles*) spanning a two-dimensional property space. Subset A, uniformly immersed in the space of the library, is visibly more representative than subset B, which differs only in two molecules located on the periphery of the data. When partitioning the chemical space by Optimum Binning, the real accessible space comprises 6 out of a total of 9 cells, so that 6 virtual points are placed into their centers (*solid squares*) following the cell-integral-diversity criterion procedure. Concerning the diversity integral criterion, a total of 27 random points (*crosses*) are uniformly distributed in the property space, occupying three unfilled bins. The normalized sums over all the minimum distances between each set of points and the molecules from each subset are as follows: 0.363 (C.I.D-subset A), 0.527 (C.I.D-subset B), 0.683 (D.I-subset A), and 0.544 (D.I-subset B). Although the absolute values of the different methods are not directly comparable, subset A is ranked above subset B by the C.I.D

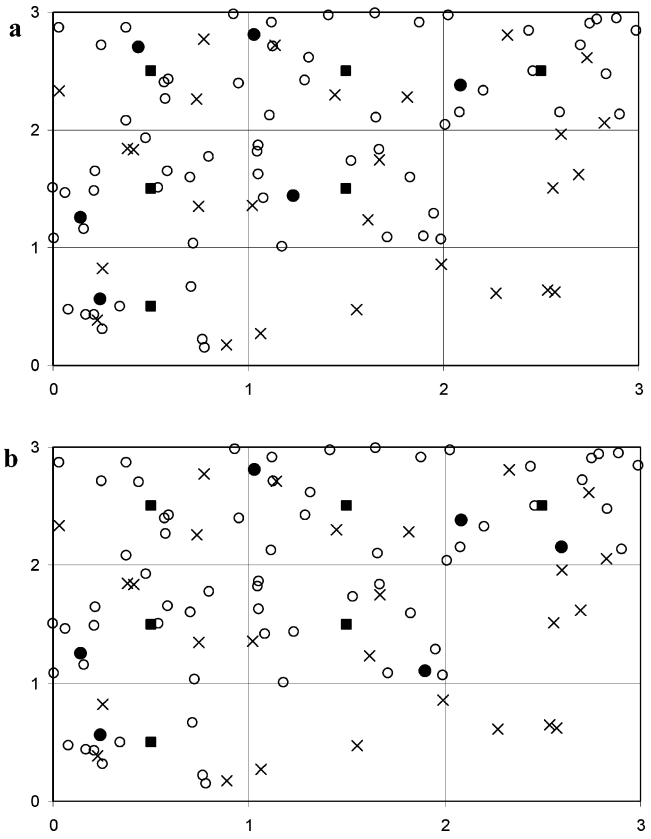

Figure 3. Influence of the actual property space sampled by the diversity integral (D.I) and the cell-integral-diversity (C.I.D) criteria. Two subsets selected (*solid dots*) from a virtual library (*unfilled circles*) are shown: (a) subset A, uniformly distributed and (b) subset B with compounds on the periphery of the data. The *solid squares* represent the C.I.D points, and the *crosses*, the D.I random points.

criterion, whereas the inverse ranking is found by the D.I criterion. This preference for compounds at the edges will depend on the particular distribution of compounds in the entire library, i.e., the ratio between the real accessible space and the hypercube defined by the diversity integral criterion.

Measurement of the Consistency of the Evaluation Methods. For all the evaluation experiments (defined by a concrete method and its settings), the three series of collections corresponding to each library were sorted in order of decreasing diversity. Then we calculated the Spearman rank-order correlation coefficient[29] (R_s, eq 4) as a measure of overall consistency in ranking the different collections between (i) an evaluation method under different settings and (ii) two different evaluation methods.

$$R_s = 1 - \frac{6\sum_{i=1}^{s} d_i^2}{s(s^2 - 1)} \tag{4}$$

In eq 4, d_i is the ranking difference of the ith selection when evaluated by two experiments (different evaluation settings or methods), and s is the number of ranked elements (21 selections from Library I and Library II and 7 selections from Library III). The Spearman correlation coefficient ranges from -1 and $+1$, where $+1$ corresponds to a perfect correlation (identical rankings), -1 corresponds to a perfect inverse correlation, and 0 corresponds to a total disorder. This coefficient has been extensively used in molecular docking studies to correlate the predicted ranking of candidate molecules with the experimental results. In our case, there is not an expectable ranking of selections to be reproduced. Here, this coefficient is merely used to analyze the influence of the different settings for a particular evaluation method (intravariation) and to establish a comparison between the methods.

RESULTS AND DISCUSSION

This section is structured as follows. The first two points deal with the analysis of the three evaluation criteria for (i) the 21 MaxMin averaged selections of different cardinality from Library I and Library II and (ii) for the 7 subsets selected by different methods from Library III. Next, the cell-integral-diversity criterion is applied in library design. Finally, details of the computational cost are given.

Collections of Different Cardinality. The three diversity measurements, SPC (a), D.I (b), and C.I.D (c), are depicted in Figures 4 and 5 for Library I and Library II, respectively. In the SPC and C.I.D graphs, the different curves correspond to each of the six binning schemes (P1...P6) taken as reference partitions (Table 2). In the case of the diversity integral method and for comparison purposes a number of random points equivalent to the occupied bins has been chosen, as it can be seen in the D.I graphs, where there is an additional seventh line for the case of fixing a total of 1000 points (Library I) or 10 000 points (Library II). The D.I values are the average of the ten repetitions for each of the seven different samplings, with their corresponding standard deviations shown in Figures 4d and 5d.

As mentioned above there is not a predictable optimal ranking of the 21 selections in order of decreasing diversity known a priori to be taken as reference. Although one would expect coverage to increase with the number of selected compounds, this is only true for cases with big differences between selection sizes or cherry picking selections. In our case the variation step between the 21 selection sizes is not significant: 1 and 5 compounds for Library I and Library II, respectively. Thus, one may find examples (as for Library I) where the array containing more compounds ($18 = 2 \times 3 \times 3$) covers less space than the subset with fewer molecules ($16 = 2 \times 2 \times 4$) when evaluated by any of the three criteria (see Figure 4a−c). Moreover, in full array designs, the number of different $n_1 \times n_2 \times ... \times n_i$ arrays derived from an $N_1 \times N_2 \times ... N_i$ i-component combinatorial library is

$$\prod_{j=1}^{i} \binom{N_j}{n_j} \tag{5}$$

Thus, the higher the number of possibilities for a concrete configuration the more probable it is to find a better representative subset. For example, for the $29 = 1 \times 1 \times 29$ and the $28 = 2 \times 2 \times 7$ arrays from Library I there are 600 and 10^8 possible subsets, respectively. Thus, as it can be seen in Figure 4a−c, the 29-membered optimal array is less diverse than the 28-membered array.

First, we analyze the intravariation of the diversity measurements using the SPC criterion (Figures 4a and 5a) under different partitions. As the number of bins of the reference partition reduces the more probable it is for a

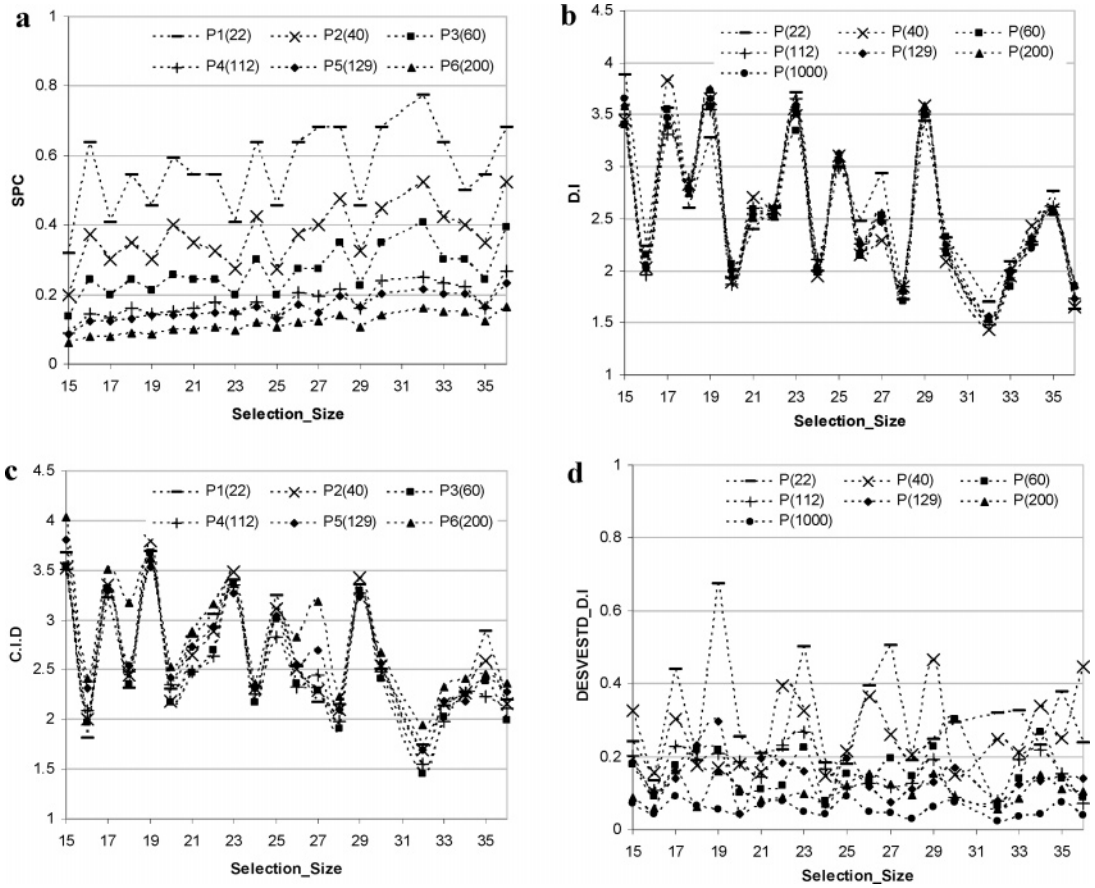

Figure 4. Diversity measurements by (a) SPC, (b) D.I (mean values), and (c) C.I.D for each of the 21 subsets of different selection size from Library I. (d) Standard deviation of the ten D.I measurements.

specific subset to achieve a greater coverage. Therefore, it is not surprising that each of the 21 selections from Library I and Library II retrieves a greater absolute value when evaluated under P1 than under P6. However, the ranking of the 21 selections in order of decreasing diversity is not conserved across the six reference partitions. For example, considering Library I in Figure 4a, the 35-membered selection is ranked above the 34-membered subset when evaluated under the P1 partition, whereas the opposite sorting is found under the rest of reference partitions (P2−P6). A similar case is observed for Library II in Figure 5a: under P1, the 250-membered subset is more diverse than the 245-membered collection, but this ranking is inverted under P5. To quantify the degree of consistency, we calculated the Spearman correlation coefficient R_s for the 15 possible pairs of the sorted lists of the 21 selections established by each of the 6 binning schemes (6*5/2). These results are given in Tables 3 and 4 for Library I and Library II, respectively (first−fifth columns). Although there is a positive correlation, the reference partition has a strong impact on the ranking of the subsets, with R_s ranging between 0.534 and 0.949 (Library I) and between 0.464 and 0.953 (Library II).

Concerning the D.I calculations in Figures 4b and 5b it can be seen that the evaluations are almost independent of the number of random points applied. Indeed, the correlation coefficients of the 21 pairwise combinations of the 7 sorted lists (7*6/2), average 0.967 (Library I) and 0.969 (Library II). We would like to emphasize that the values shown in Figures 4b and 5b are the mean of ten repetitions. When a unique diversity evaluation run is considered, a certain degree of agreement between the seven curves is lost and R_s decreases. In the case of Library II, the R_s mean value of the 21 crossed-rankings changes from 0.969 (10 repetitions) to 0.815 (1 repetition). Concerning the magnitude of the standard deviation it depends both on the number of sampling probes (the greater the number, the lower the standard deviation) and on the particular case, that is the library under study (compare Figures 4d and 5d).

Going on to the C.I.D graphs, the diversity profiles are very similar to those obtained for the D.I index. In Tables 5 and 6 the correlation coefficients for all 15 pairwise combinations of the rankings obtained for the six reference partitions are tabulated (first−fifth columns) for Library I and Library II, respectively. For both libraries, the six lists of ranked selections are in good agreement, with R_s ranging between 0.821 and 0.992 (Library I) and between 0.812 and 0.984 (Library II). Compared to the lower R_s values obtained for the cell-based method, the influence of the partition is greatly reduced with the cell-integral-diversity criterion. We mean the classification/sorting of the set of selections based on the cell-integral-diversity criterion within the different partitions of reference does not vary as much as it does with the cell-based method. This difference can be attributed to the avoidance of the cluster artifact (Figure 2) achieved by this new method, being other factors such as the intrinsic distribution of compounds into the bins equivalent for both cases.

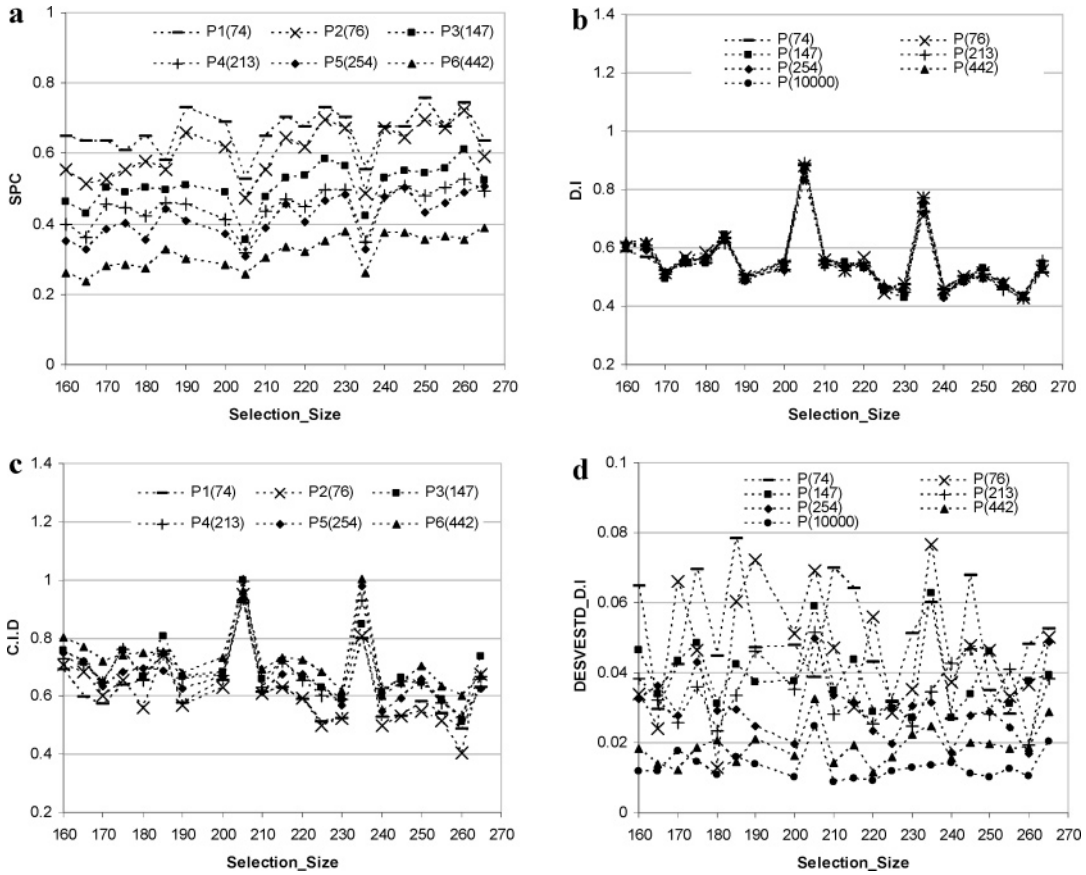

Figure 5. Diversity measurements by (a) SPC, (b) D.I (mean values), and (c) C.I.D for each of the 21 subsets of different selection size from Library II. (d) Standard deviation of the ten D.I measurements.

Table 3. Spearman Rank Correlation for the Ranking of the 21 Selections from Library I by the Six Cell-Based Measurements and the Diversity Integral Criterion with 1000 Points

R_s values	SPC (P1)	SPC (P2)	SPC (P3)	SPC (P4)	SPC (P5)	D.I ($p = 1000$)
SPC (P1)						0.842
SPC (P2)	0.922					0.944
SPC (P3)	0.845	0.949				0.918
SPC (P4)	0.590	0.673	0.730			0.753
SPC (P5)	0.534	0.645	0.657	0.894		0.641
SPC (P6)	0.705	0.801	0.861	0.677	0.653	0.734

Table 4. Spearman Rank Correlation for the Ranking of the 21 Selections from Library II by the Six Cell-Based Measurements and the Diversity Integral Criterion with 10 000 Points

R_s values	SPC (P1)	SPC (P2)	SPC (P3)	SPC (P4)	SPC (P5)	D.I ($p = 10\,000$)
SPC (P1)						0.731
SPC (P2)	0.887					0.877
SPC (P3)	0.722	0.887				0.882
SPC (P4)	0.539	0.791	0.927			0.861
SPC (P5)	0.477	0.743	0.844	0.947		0.795
SPC (P6)	0.464	0.739	0.809	0.894	0.953	0.775

Table 5. Spearman Rank Correlation for the Ranking of the 21 Selections from Library I by the Six Cell-Integral-Diversity Measurements and the Diversity Integral Criterion with 1000 Points

R_s values	C.I.D (P1)	C.I.D (P2)	C.I.D (P3)	C.I.D (P4)	C.I.D (P5)	D.I ($p = 1000$)
C.I.D (P1)						0.897
C.I.D (P2)	0.981					0.936
C.I.D (P3)	0.961	0.992				0.945
C.I.D (P4)	0.901	0.935	0.955			0.914
C.I.D (P5)	0.887	0.925	0.945	0.969		0.916
C.I.D (P6)	0.821	0.879	0.906	0.948	0.958	0.938

Table 6. Spearman Rank Correlation for the Ranking of the 21 Selections from Library II by the Six Cell-Integral-Diversity Measurements and the Diversity Integral Criterion with 10 000 Points

R_s values	C.I.D (P1)	C.I.D (P2)	C.I.D (P3)	C.I.D (P4)	C.I.D (P5)	D.I ($p = 10\,000$)
C.I.D (P1)						0.926
C.I.D (P2)	0.903					0.923
C.I.D (P3)	0.899	0.927				0.868
C.I.D (P4)	0.892	0.917	0.983			0.869
C.I.D (P5)	0.847	0.812	0.818	0.855		0.904
C.I.D (P6)	0.873	0.843	0.826	0.862	0.984	0.935

After judging the intravariation of each method with regard to different intrinsic evaluation conditions (partitions or number of points), we compared the degree of agreement between the three diversity indices.

Based on the coherence across the seven determinations we took the ranking of the 21 collections obtained by the D.I index with 1000 points (Library I) or 10 000 points (Library II) as reference. In the last column of Tables 3−6 (labeled "D.I") we summarize the corresponding correlation coefficients between the sorted full arrays by the D.I index ($p = 1000$ or $p = 10000$) and each of the six measurements carried out with the cell-based (Tables 3 and 4) and the cell-integral-diversity (Tables 5 and 6) methods. For both libraries, the D.I correlates stronger with the C.I.D rankings than with the SPC sorted lists as it would be expected given their modes of operation. For Library I, the mean of the correlation coefficients for the D.I−C.I.D and the D.I−SPC

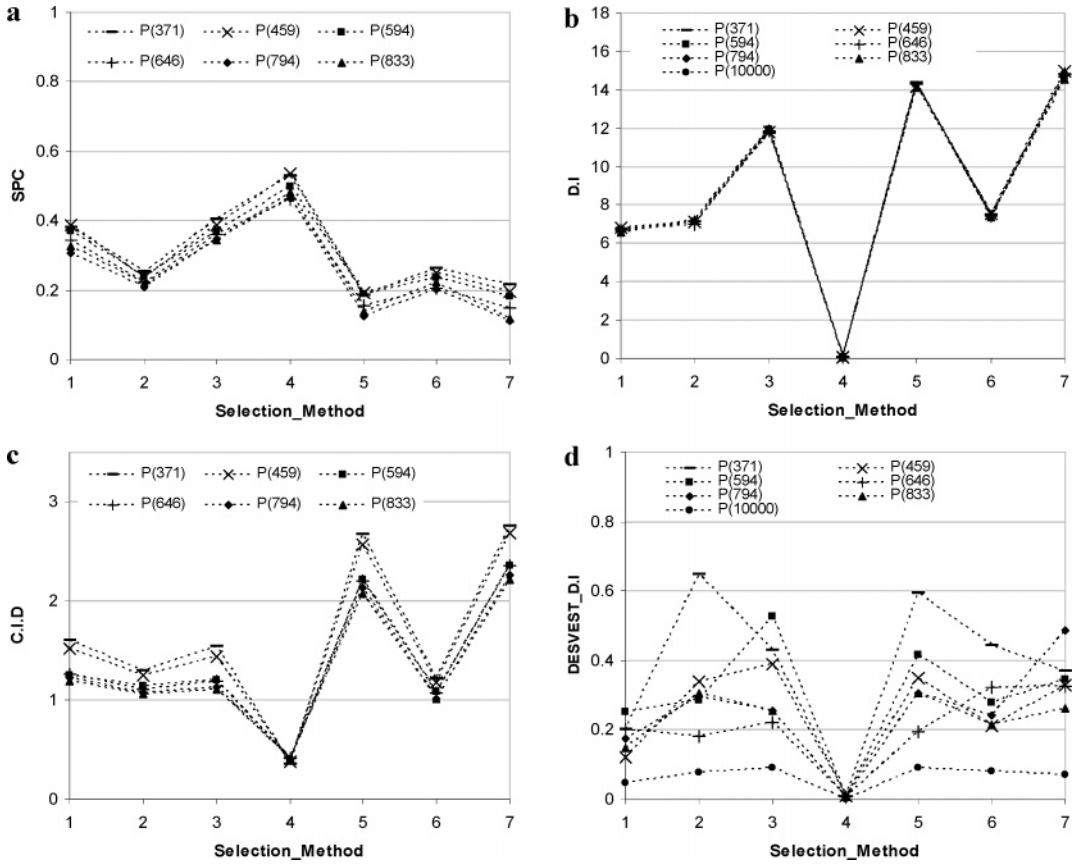

Figure 6. Diversity measurements by (a) SPC, (b) D.I (mean values), and (c) C.I.D for each of the 7 subsets selected by different methods from Library III. (d) Standard deviation of the ten D.I measurements. 1: MaxMin full array; 2: MaxMin sparse array; 3: Optimum Binning full array; 4: Optimum Binning sparse array; 5: K-means full array; 6: K-means sparse array; 7: Random.

Table 7. Ranking of the Seven Subsets Selected by Different Methodologies from Library III in Order or Decreasing Diversity as Calculated by the Three Criteria: C.I.D, D.I, and SPC

			SPC	
diversity index evaluation setting	C.I.D P1−P6	D.I all sets of points	P1−P2	P3−P6
rank 1	Binning Sparse	Binning Sparse	Binning Sparse	Binning Sparse
rank 2	K-means Sparse	MaxMin Full	Binning Full	Binning Full
rank 3	MaxMin Sparse	MaxMin Sparse	MaxMin Full	MaxMin Full
rank 4	Binning Full	K-means Sparse	MaxMin Sparse	MaxMin Sparse
rank 5	MaxMin Full	Binning Full	K-means Sparse	K-means Sparse
rank 6	K-means Full	K-means Full	Random	K-means Full
rank 7	Random	Random	K-means Full	Random

pairs are 0.924 and 0.805, respectively, whereas for Library II these values are 0.904 and 0.820. In both libraries, the average correlation coefficient of the D.I−C.I.D pairs is very similar to the values obtained for the intravariation of the D.I method, where the small difference may be attributed to the differences in the actual space sampled by the two evaluation methods (Figure 3).

Collections Selected by Different Methods. In Figure 6 the three diversity marks (SPC (a), D.I (b), and C.I.D (c)) of the selections from Library III obtained by 7 different selection methods (MaxMin Sparse, MaxMin full array, Optimum Binning Sparse, Optimum Binning full array, K-means Sparse, K-means full array, and Random) are depicted. The rankings of the methods within C.I.D (Figure 6c) and D.I (Figure 6b) criteria remain unchanged when altering the evaluation settings (that is the number of partitions and the number of random points, respectively), yielding in all cases $R_s = 1$. With regard to the cell-based criterion, the variation of the reference partition gives rise to two nearly equal sorting of the methods according to space coverage, differing only in the rank position between the K-means full array and the random selections. In Table 7, we summarize the four possible rankings of the seven methods found by the three diversity indices. Within each list the methods appear in order of decreasing diversity. The sparse binning selection is clearly ranked as the most diverse selection by the three diversity indices. The K-means full array selection is almost indistinguishable from the random selection by any of the three methods (Figure 6a−c), so its interchange with the random selection depending on the reference partition of the cell-based evaluation criterion is not a major drawback. To summarize the extreme cases, that is, considering sparse binning selections as the most diverse and K-means full arrays as the least coincide for all three evaluation criteria.

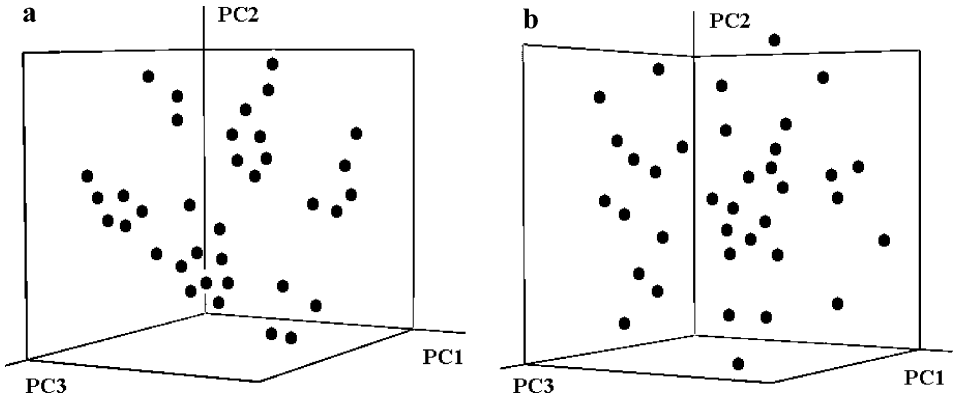

Figure 7. Full arrays selections from Library I shown in the space of the first three principal components (69% of the variance): (a) 3 × 2 × 6 selection that maximizes the cell occupancy and (b) 3 × 2 × 6 array that minimizes the cell-integral-diversity metric.

It should be noted that the cell-based method prioritizes those selections carried out by the Binning algorithm, as both the selection and the evaluation criteria coincide. The two other criteria, which measure distances instead of occupied cells, rank the full array binning selection as less diverse in front of other distance-based selections including even the MaxMin full array selection as well as cluster based selections represented by K-means sparse. Concentrating on the MaxMin selected subsets, we have observed that some compounds in the edges of the descriptor space are responsible for the D.I score of those collections following the idea depicted in Figure 3. Therefore, these subsets appear at preferential rank positions in the D.I ranking. However, it should be considered that the D.I marks of the K-means sparse, MaxMin sparse, and MaxMin full array selections are very similar, with overlapping standard deviations (Figure 6b,d). The same applies for the C.I.D index (Figure 6c). However, in this case, it is remarkable that the method ranks sparse selections above the corresponding full array designs, an expected tendency that is not observed by the other evaluation metrics, as the combinatorial restriction implies a reduction of coverage.

Cell-Integral-Diversity Criterion in Library Design. After analyzing the ability of the cell-integral-diversity criterion to evaluate and assess diversity, we examined the use of the C.I.D metric in the context of library design, particularly of full array library selections, where it can be extremely useful as full arrays group the compounds in space in such a way that cluster artifacts in cell-based metrics occur frequently.

Briefly its working scheme for full array selection consists in an Optimum Binning partition of space followed by placing the probes in the center of the occupied bins and finally minimizing the objective function given in eq 6 by the previously described genetic algorithm

$$\text{C.I.D} = \frac{\sum_{i=1}^{p} \min_{j \subset n} D_{ij}}{p} \qquad (6)$$

where p corresponds to the total number of points. Three selections were carried out from each Library (I, II and III), aiming at identifying the best 3 × 2 × 6, 19 × 10, and 6 × 6 × 6 × 6 combinatorial subsets, respectively, using the C.I.D metric (eq 6). For evaluation purposes, the same selections were optimized using the cell fraction (SPC) metric. For Library I the used partitioning algorithm generates 22 occupied bins (P1 in Table 2), whereas 147 (P3 in Table 2) and up to 1209 filled bins are obtained for the 35 640-compound and the 531 441-compound libraries, respectively.

In Figure 7 the resulting SPC selection (7a) and C.I.D selection (7b) for Library I have been depicted in 3 dimensions that correspond to the first three PCs, which account for 69% of the variance. It can be observed that the SPC selection, which covers 21 out of 22 occupied bins, contains groups of compounds very similar to each other, whereas the C.I.D selection, which spans fewer bins (17), contains molecules spaced further apart exploring uniformly the space.

The same conclusions can be drawn from the two selections from Library II, highlighted in parts a (SPC selection) and b (C.I.D selection) of Figure 8. For clarity, we depict here the 2-D plots based on the first two PCs, representing 63% of the variance. In this case the cell-based and the cell-integral-diversity selections occupy 109 and 82 out of 147 total bins. Despite the lesser degree of coverage of the cell-integral-diversity selection in terms of occupied cells, this reduction is not visually observed in terms of distances. As in the previous case, the C.I.D subset is more homogenously distributed within the bins than the SPC selection, where compounds tend to group in the central region. The SPC selection gives a C.I.D score of 3.816, whereas the C.I.D score for the library optimized with the cell-integral-diversity function is 3.507. (As reference, the minimum C.I.D score for the parent library is 3.107.)

Finally, in parts a (SPC) and b (C.I.D selection) of Figure 9 the 2-D plots based on the first two PCs (52% of the variance) are given for the 6 × 6 × 6 × 6 combinatorial selections from Library III. Here, the SPC optimized selection covers 423 out of 1209 occupied cells, and the C.I.D selection spans 359 cells. The C.I.D marks for each selection are 2.446 (SPC) and 2.325 (C.I.D-optimized library), compared to the total mark of 1.648 for the entire library of 531 441 compounds. On the basis of the visual inspection, the C.I.D function manages to find a library that covers the space better than the selection optimized using the SPC metric.

Computational Cost. As stated above the advantages of C.I.D over D.I are its deterministic character and reduced time complexity. In that sense we have registered the computational effort for a series of calculations depending

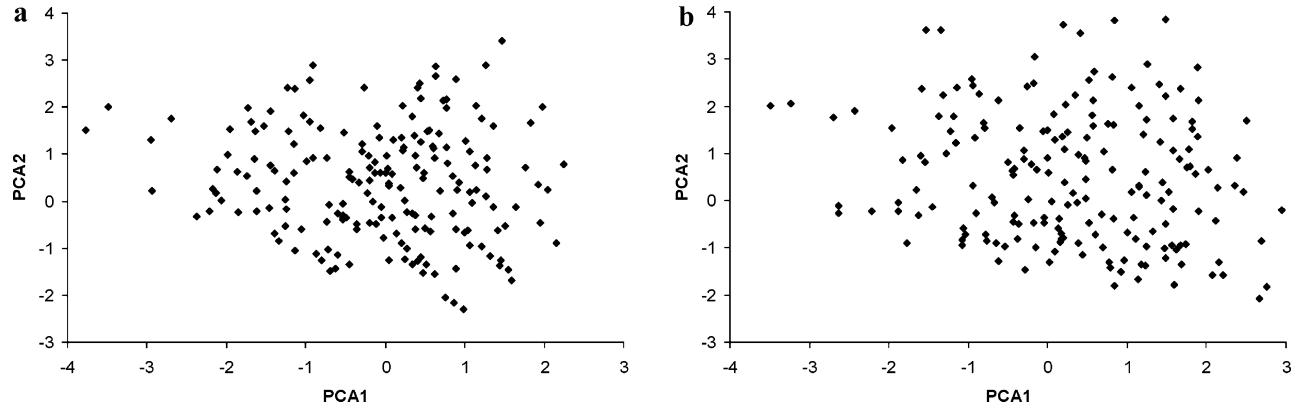

Figure 8. Sublibraries from Library II shown in the space of the first two principal components (63% of the variance): (a) 19 × 10 selection that maximizes the cell occupancy and (b) 19 × 10 array that minimizes the cell-integral-diversity metric.

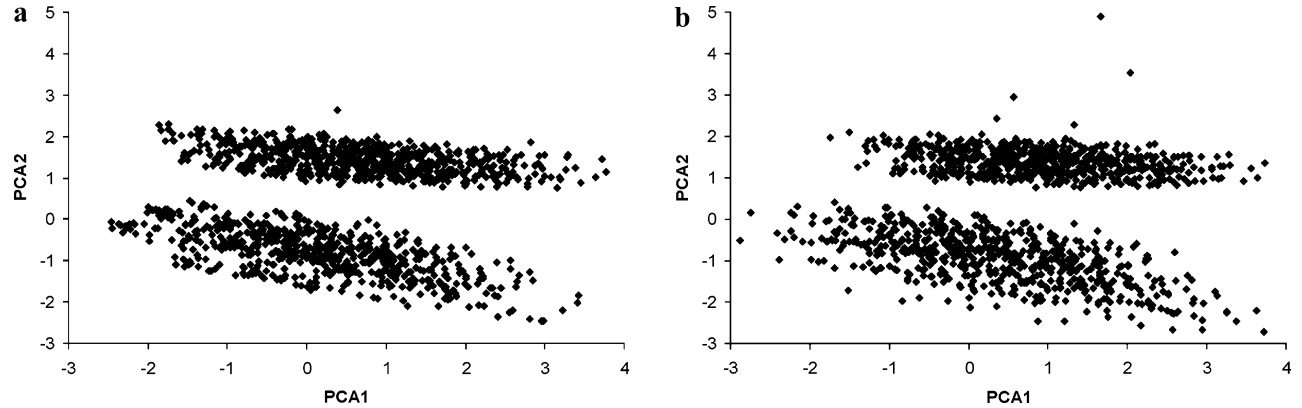

Figure 9. Sublibraries from Library III shown in the space of the first two principal components (52% of the variance): (a) 6 × 6 × 6 × 6 selection that maximizes the cell occupancy and (b) 6 × 6 × 6 × 6 array that minimizes the cell-integral-diversity metric.

Table 8. Timings for the Three Diversity Indices When Evaluating Subsets of Different Cardinality from Library I, Library II, and Library III under Different Evaluation Settings

	Library I ($N = 600$)			Library II ($N = 35640$)			Library III ($N = 531441$)	
		$n = 15$	$n = 36$		$n = 160$	$n = 265$		$n = 1296$
SPC	P1 (22 bins)	<0.01	<0.01	P1 (74 bins)	0.06	0.07	P1 (371 bins)	11.36
	P6 (200 bins)	0.01	0.01	P6 (442 bins)	0.13	0.14	P6 (833 bins)	16.47
D.I	22 points	0.01	0.01	74 points	1.20	1.21	371 points	74.21
	200 points	0.04	0.05	442 points	7.00	7.04	833 points	169.93
	1000 points	0.23	0.24	10000 points	158.62	168.10	10000 points	1987.04
C.I.D	P1 ($p = 22$)	0.01	0.01	P1 ($p = 74$)	1.24	1.27	P1 ($p = 371$)	83.14
	P6 ($p = 200$)	0.04	0.05	P6 ($p = 442$)	7.13	7.21	P6 ($p = 833$)	180.90

on the size (n) of the array being evaluated and the evaluation settings selected for each method (Table 8 indicates the CPU usage in seconds.). All the calculations were run on a single 2.4 GHz Pentium IV Linux workstation. The time needed to carry out the selection is not included. The computational cost associated with the Binning algorithm employed by the C.I.D method is negligible compared to the distance calculation effort, and therefore a single C.I.D run is comparable to a single D.I run in terms of efficiency. In both cases, the CPU usage of the methods correlates linearly with the number of points. However, one should keep in mind that a number of repetitions are usually required for D.I in order to attain a stable profile as it is the case for the calculations shown in Figures 4b, 5b, and 6b.

On the other hand, the timings (CPU usage) required to optimize the 19 × 10 and 6 × 6 × 6 × 6 selections from Library II and Library III are listed in Table 9. It can be seen that the cell-integral-diversity optimizations run slower than the cell fraction calculations, as the former requires the determination of the nearest neighbor compound for each point of the binning scheme.

Table 9. Timings for Optimizing the Selections from Library II and Library III by Using the Cell Fraction Criterion and the C.I.D Criterion with a Genetic Algorithm[a]

	Library II $n = 190$	Library III $n = 1296$
SPC	7.23 s (2400 iterations)	6.60 min (2188 iterations)
C.I.D	18.12 min (2775 iterations)	3.90 h (2215 iterations)

[a] In all cases a population of 30 individuals was used, so the number of function evaluations corresponds to the product of the number of individuals and the number of iterations.

CONCLUSIONS

The cell-based and the diversity integral criterion are two techniques traditionally applied to assess and compare the diversity of collections defined in terms of space coverage. We have analyzed the influence of their parameters (refer-

ence partition and number of points, respectively) on their efficacy. Partition techniques show a great dependency on the chosen partition reference, so that their ability to rank sets of collections in order of decreasing diversity is affected by the arbitrariness of the cell boundaries. The cell-integral-diversity method has proven to be very useful to bypass cluster artifacts, obtaining higher Spearman rank-order correlation coefficients between diversity evaluations carried out with different partition references on different sized libraries. It constitutes a valid option for the diversity assessment of collections having different cardinalities and properly discerns between subsets identified by rational selection methods and random selections. Furthermore, it avoids the heuristic character of the diversity integral criterion and therefore the need to repeat calculations, reducing in that way the global computational cost. This is particularly useful in the context of large libraries.

From the point of view of full array design, the cell-integral-diversity metric is a good alternative to the commonly used cell fraction metric despite the increase in computational cost. The use of distances instead of cell counts circumvents the cluster artifact achieving thereby higher diversity. Despite reducing the number of sampled cells, the cell-integral-diversity optimized selections are visually regarded as more diverse.

It should be mentioned that these methods are only aimed at optimizing the diversity with the underlying assumption that a diverse collection will maximize the likelihood of finding hits by testing as many chemically distinct compounds as possible and avoiding redundancies in screening experiments.

ACKNOWLEDGMENT

One of us (O.R.) would like to thank the Generalitat de Catalunya - DURSI for a grant within the Formació de Personal Investigador (2003FI) program. Support of this work by a grant from The TV3 Marathon Foundation (AIDS-2001) promoted by the Catalan Radio and Television Corporation [Corporació Catalana de Ràdio i Televisió] (CCRTV) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Johnson, M. A.; Maggiora, G. M. Introduction of similarity in Chemistry. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990; pp 1−13.

(2) Willett, P. Subset-Selection Methods For Chemical Databases. In *Molecular Diversity in Drug Design*; Dean, P. M., Lewis, A. R., Eds.; Kluwer Academic Publishers: The Netherlands, 1999; pp 115−140.

(3) Perez, J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34*, 143−152.

(4) Pearlman, R. S.; Smith, K. M.; Deanda, F. Low-dimensional chemistry spaces: recent advances. Paper presented at the Cambridge Healthtech Institute conference "Chemoinformatics" held in Boston June 15−16, 1998.

(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(6) Lobanov, V. S.; Agrafiotis, D. K. Stochastic Similarity Selections from Large Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 460−470.

(7) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64−74.

(8) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: a flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60−66.

(9) Waldman, M.; Li, H.; Hassan, M. Novel algorithms for the optimization of molecular diversity of combinatorial libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 412−26, 533−6.

(10) Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.

(11) Jamois, E. A.; Hassan, M.; Waldman, M. Evaluation of Reagent-Based and Product-Based Strategies in the Design of Combinatorial Library Subsets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 63−70.

(12) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*: Lipkowitz, K. B., Boyd, B. D., Eds.; VCH: New York, 2002; Vol. 18, pp 1−40.

(13) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(14) Gillet, V. J.; Willett, P.; Bradshaw, J. The Effectiveness of Reactant Pools for Generating Structurally-Diverse Combinatorial Libraries. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 731−740.

(15) Zheng, W. Z.; Waller, C. L.; Cho, S. J.; Tropsha, A. Rational Combinatorial Library Design. 3. Simulated Annealing Guided Evaluation (SAGE) of Molecular Diversity: a Novel Computational Tool for Universal Library Design and Database Mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738−746.

(16) Agrafiotis, D. K. A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 159−167.

(17) Reynolds, C. H.; Tropsha, A.; Pfahler, L. B.; Druker, R.; Chakravorty, S.; Ethiraj, G.; Zheng, W. Diversity and coverage of structural sublibraries selected using the SAGE and SCA algorithms. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1470−1477.

(18) Graham, E. T.; Jacober, S. P.; Cardozo, M. G. A Novel Frequency Distribution Selection Method for Efficient Plate Layout of a Diverse Combinatorial Library. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1508−1516.

(19) *Cerius2, version 4.6*; Molecular Simulations Inc.: 9685 Scranton Rd., San Diego, CA 92121, 2001.

(20) Pascual, R.; Borrell, J. I.; Teixidó, J. Analysis of selection methodologies for combinatorial library design. *Mol. Diversity* **2003**, *6*, 121−133.

(21) Maloney, P. R.; Parks, D. J.; Haffner, C. D.; Fivush, A. M.; Chandra, G.; Plunket, K. D.; Creech, K. L.; Moore, L. B.; Wilson, J. G.; Lewis, M. C.; Jones, S. A.; Willson, T. M. Identification of a chemical tool for the Orphan Nuclear Receptor FXR. *J. Med. Chem.* **2000**, *43*, 2971−2974.

(22) Mont, N.; Teixidó, J.; Borrell, J. I.; Kappe, O. A three-component synthesis of pyrido[2,3-*d*]pyrimidines. *Tetrahedron Lett.* **2003**, *44*, 5385−5387.

(23) *MOE (Molecular Operating Environment*), *2004.03 Release*; Chemical Computing Group Inc.: 1010 Sherbrooke West, Suite 910, Montreal H3A 2R7, Canada, 2004.

(24) Halgren, T. A. The Merck Force Field. *J. Comput. Chem.* **1996**, *17* (5), 490−519.

(25) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.* **2000**, *18*, 464−477.

(26) Brown, R.; Martin, Y. C. Designing Combinatorial Library Mixtures Using a Genetic Algorithm. *J. Med. Chem.* **1997**, *40*, 2304−2313.

(27) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(28) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics Modell.* **1997**, *15*, 372−385.

(29) Lane, D. M. HyperStat Online: An Introductory Statistics Textbook and Online Tutorial for Help in Statistic. http://www.davidmlane.com/hyperstat/index.html (accessed May 30, 2006).

CI600433C