Data Structure Comparison Using Box Counting Analysis

Yukio Tominaga†

Department of Chemistry I, Discovery Research Laboratories I Dainippon Pharmaceutical Co., Ltd. Enoki 33-94, Suita, Osaka 564, Japan

Received April 20, 1998

Box counting analysis was performed to visualize complex data structures of datasets. Two datasets were used as original datasets. One included 8000 samples and the other included 53 064 samples. Nine different selection methods were used to select subsets. The selection methods were as follows: maximum dissimilarity method, maximum similarity method, group averaging hierarchical clustering method, reciprocal nearest neighbor Ward hierarchical clustering method, k-mean nonhierarchical clustering methods with two types of seed points, Genetic algorithms (GAs), Kohonen networks, and cell-based method. The data structures of selected subsets were compared to that of the original dataset by using box counting analysis.

INTRODUCTION

Since the concept of molecular diversity was introduced into the field of medicinal chemistry,[1,2] the concept has been applied to drug discovery programs, especially to lead compound finding process.[3−7] When biological screening was performed to find lead compounds, all the compounds available were not used, but selected subsets were used to save time and money. The subsets include almost equivalent information to all the compounds available. Therefore, there has been a great deal of interest to select a representative subset from all the compounds available. To select a subset, one usually assigns chemical and physical descriptors to each compound first, then compounds are classified by these descriptors, and finally compounds are selected from each class as a subset.[8−10] The aim of this selection is to generate as diverse a subset as possible and reduce redundancy of an original dataset.

Three principal methods have been applied to select subsets: (I) cluster-based selection; (II) dissimilarity-based selection; and (III) cell-based selection.[11−14] Cluster-based method can be divided into hierarchical and nonhierarchical methods. Hierarchical methods, for example, group averaging method and Ward method, classify compounds in an ascending or descending manner by constructing treelike dendograms. Nonhierarchical methods are based on the optimization of an objective function. In the methods there are no parent−daughter relationships between the various clusters. There are many types of nonhierarchical clustering methods such as leader algorithms, k-mean method, Jarvis−Patric method, Kohonen self-organizing feature map, and so on. Dissimilarity-based method identifies the set of most dissimilar molecules in a dataset using some quantitative measure of dissimilarity. In the cell-based method, multidimensional descriptor space is divided into uniform hyperboxes called cells, and a molecule is selected from each cell to select a diverse subset.

The comparative studies for these selection procedure are very important to identify the best selection procedure. In our previous study, the complex data structures of the randomly selected subset and the representative subset which was selected by nonhierarchical clustering method were compared and successfully distinguished by using box counting analysis.[15,16] Box counting analysis is widely used to determine fractal dimensions of complex structures.

In this study, further comparative studies were performed from the viewpoint of the data structure. Two datasets were used as original datasets. The dataset I with 8000 samples was a moderate size dataset, and the dataset II with 53 064 samples was a relatively large size dataset. The reason we used the dataset I with moderate size in this study is that we wanted to compare the data structures of the subsets selected with as many selection methods as possible. Some selection methods are not appropriate to apply a large dataset such as dataset II. Nine different kinds of subset selection methods were applied to dataset I to select subsets with 700 samples. The subset selection methods were as follows: (1) maximum dissimilarity method; (2) maximum similarity method;[17,18] (3, 4) two types of hierarchical clustering methods, group averaging hierarchical clustering method and reciprocal nearest neighbor Ward hierarchical clustering method;[19−20] (5, 6) k-mean nonhierarchical clustering methods[19] with two types of seed points [Some seed points were selected randomly, and the other seed points were selected by leader algorithm.];[19] (7) Genetic algorithms (GAs)[21−23] [The fitness function of GAs was clustering efficiency when samples of a subset were set to seed points of k-mean nonhierarchical clustering method]; (8) Kohonen networks;[24−27] and (9) cell-based method. The cell-based method and the maximum similarity method were used to know the characteristics of the data structures of diverse and redundant dataset, respectively. As for the relatively large dataset II, five different selection methods which could process a large dataset were applied to select the subsets with 1000 samples. The selection methods were reciprocal nearest neighbor Ward hierarchical clustering method, two types of k-mean nonhierarchical clustering methods, Kohonen networks, and cell-based method. The data structures of the selected subsets were compared by using box counting analysis.

† Tel.: +81-6 337-5898. Fax.: +81-6 338-7656. E-mail: yukio-tominaga@dainippon-pharm.co.jp.

S0095-2338(98)00207-8 CCC: $15.00
Published on Web 08/08/1998

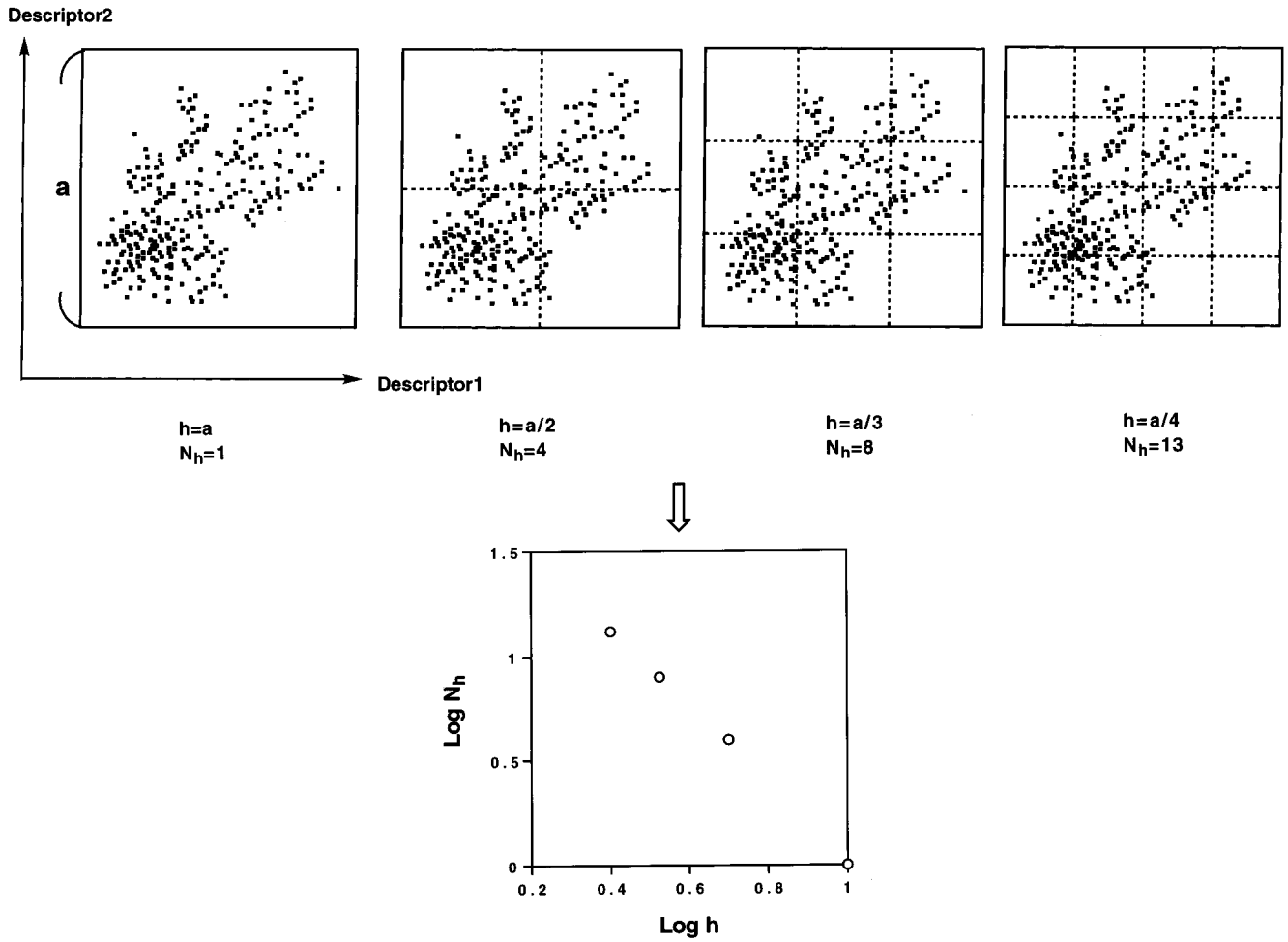

Figure 1. Schematic representation of box counting plots.

METHOD

1. Box Counting Analysis. Box counting analysis is widely used[15,16] to estimate fractal dimension of complex data pattern. Data points are overlaid by a uniform grid with given grid spacing, *h*, in such manner that the minimum number of boxes can cover the full data points. The number of boxes (N_h) containing any data points are counted. A grid with reduced *h* is then laid over the data points, and the counting process is repeated with different *h*. N_h is plotted against *h* on logarithmic scale. This process is schematically represented in Figure 1. The plot contains the information about the data structure, e.g., the absolute value of the gradient is the fractal dimension, D_f, when there is a linear relation between log h and log N_h. The fractal dimension is mathematically expressed as

$$D_f = \lim_{h \to 0} (\log N_h)/(\log(1/h)) \qquad (1)$$

where D_f is the fractal dimension, *h* is the size of measuring elements, and N_h is the number of measuring elements of size *h*, which is required to approximate the structure. Now to compare the data structures of selected subsets and an original dataset, the deviation between log N_h of an original dataset and that of each subset (dlog N_h) was estimated.

2. Subset Selection Methods. 2.1. Maximum Dissimilarity Method and Maximum Similarity Method. Every newly selected sample is maximally (dis)similar from the previously selected samples. Euclidean distance was used to estimate (dis)similarity. If the summation of Euclidean distance between a sample and the previously selected samples is smallest (largest), the sample was selected. The selection stops when the number of selected samples reaches user specified number. The algorithm is as follows.

Step 1: The geometrical gravity of all the samples are evaluated. The sample which locates the nearest position to the gravity is selected.

Step 2: The summation of Euclidean distances between a sample and previously selected samples is evaluated. If the summation is smallest (largest) among all the samples except previously selected samples, the sample is selected.

Step 3: Step 2 is repeated until the number of selected samples reaches user specified number.

2.2. Hierarchical Clustering Method. Group averaging hierarchical clustering method and reciprocal nearest neighbor Ward hierarchical clustering method were performed by using Tsar V.3.1.[28]

2.3. k-Mean Nonhierarchical Clustering Method. k-Mean nonhierarchical clustering methods were performed by using two types of seed points. One type of seed points was selected randomly, and the other type of seed points was selected by leader algorithm.

The leader algorithm is implemented as described below.

Step 1: Set CN as the number of cluster leaders.

Step 2: Set T as threshold of the leader algorithm.

Step 3: Set the first cluster leader as compound 1.

Step 4: When the Euclidean distance between compound 2 and the first cluster leader is less than or equal to T, compound 2 belongs to the first cluster. When the Euclidean distance between compound 2 and the first cluster leader is greater than T, compound 2 is set as second cluster leader. Repeat this for subsequential compounds.

Step 5: When the number of cluster leaders is less than CN, reduce T, return to Step 2, and repeat Steps 3 and 4. When the number of cluster leaders is more than CN, increase T, return to Step 2, and repeat Steps 3 and 4.

The k-mean clustering algorithm is implemented as described below.

Step 1: Assign each compound to the cluster containing the nearest seed point.

Step 2: Evaluate the mean of each cluster.

Step 3: Assign each compound to the cluster containing the nearest mean.

Step 4: Find new seed compounds which locate nearest to the mean of each cluster.

Step 5: If new seed points were different from the seed points in Step 1, return to Step 1.

Once clustering is completed, each cluster center is selected as a subset.

2.4. Generic Algorithms (GAs). Prior to explanation of the GAs itself, we introduce the evaluation procedure of the effectiveness of the clustering method. When compound *I* is given in cluster *J*, the predicted value of each descriptor for *I* can be estimated as the mean of the values of each descriptor for all the other compounds in cluster *J*. This procedure is repeated for each of the *N* compounds in the dataset, then the correlation is established between the sets of *N* predicted and *N* observed values of each descriptor. The correlation between the sets of observed (x) and predicted (y) values is calculated by the product-moment correlation coefficient (PMCC) denoted by r.[29] This is given by

$$r = \sum(x - \bar{x})(y - \bar{y})/\{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2\}^{1/2} \quad (2)$$

where $\bar{x}$ and $\bar{y}$ are the mean of observed and predicted values, respectively, and where the summations are all the samples occurring in clusters containing at least three members. Here we used the mean of the PMCCs of individual descriptors (MP) to evaluate the effectiveness of clustering.

$$\mathrm{MP} = (1/P) \sum_{I}^{\text{all variables}} r_{\mathrm{i}} \quad (3)$$

The algorithm of GAs is implemented as described below.

2.4.1. Coding. The ID number is assigned to each sample to select the samples for subsets as a nonbinary string. A set of strings forms a population.

2.4.2. Fitness Function. The fitness function is MP values obtained when the samples in a subset are set to the seed points of k-mean methods.

Step 1: Initial Population. A population of 50 different random combinations of strings are generated. k-Mean clustering (Steps 1−4 in the previous k-mean clustering algorithm) are performed by using the strings as the seed points. MP values are evaluated. The strings are replaced with the ID numbers of the samples which are nearest to each cluster center. Ten strings with highest fitness values are then selected as the initial population.

Step 2: Selection. Three pairs of strings (parents) are selected by using roulette wheel selection method, a procedure similar to spinning a roulette wheel with each member of the string having a portion of the area of the wheel that is proportional to its fitness.

Step 3: Crossover. Three pairs of parents are crossovered by using steady state without duplicates technique. Because strings are not in a sequential order, using a crossover operation to generate new strings may result in having strings with duplicated ID numbers.

Step 4: Mutation. Each offspring is subject to random single-point mutation; an ID number is replaced with another while avoiding any duplicates.

Step 5: Evaluation. Using the method identical to Step 1, k-mean clustering is carried out for the generated offsprings.

Step 6: Exploration. When the highest fitness value (MP) of an offspring is superior to the lowest fitness value of the initial population, the offspring is replaced with the parent.

Steps 2−6 are repeated 50 times.

2.5. Kohonen Networks. Kohonen networks can be used for the projection of multidimensional data into two-dimensional topological space. The algorithm was as follows.

Step 1: Weight vector for each neuron is set to random value. The length of each weight vector is set to 1.0.

$$\sum_{i=1}^{\text{all variables}} w_{ji} = 1.0 \quad (4)$$

where j is a ID number of neuron.

Step 2: A sample is selected at random from the original dataset. The vector of descriptors for the sample is compared with each of the weight vectors by calculating Euclidean distance

$$d_{\mathrm{sj}} = [\sum_{i=1}^{m}(x_{\mathrm{si}} - w_{\mathrm{ji}})^2]^{1/2} \quad (5)$$

where s is a ID number of selected sample. The neuron for which d_{sj} is a minimum is selected as the winning neuron, c.

Step 3: The weight vectors are modified according to the following learning rule:

$$\Delta w_{\mathrm{ij}} = \eta(t)a(d_{\mathrm{c}} - d_{\mathrm{j}})(x_{\mathrm{i}} - w_{\mathrm{ij}}) \quad (6)$$

T is the number of samples entered into the training process since the beginning of the training. $D_{\mathrm{c}}-d_{\mathrm{j}}$ is the topological distance between neuron j and wining neuron c in two-dimensional array with toroidal topology. To count $d_{\mathrm{c}}-d_{\mathrm{j}}$, hexagonal neighborhood is used. $\eta(t)$ controls the learning rate. It is a linear, decreasing function of t. The triangular function is used as the neighborhood function, $a(d_{\mathrm{c}}-d_{\mathrm{j}})$, which decreases with the topological distance, $d_{\mathrm{c}}-d_{\mathrm{j}}$.

Step 4: Steps 2 and 3 are repeated 200 000 times.

All the samples are mapped into the two-dimensional topological map. The geometrical gravity of the samples for each neuron is estimated when samples are assigned. The

Table 1. Descriptors in This Study

no.	descriptor	no.	descriptor
1	molecular mass	31	Kier chiV4 (path/cluster) index
2	molecular volume	32	Kier chi3 (path) index
3	inertia moment 1 size	33	Kier chi4 (path) index
4	inertia moment 2 size	34	Kier chi5 (path) index
5	inertia moment 3 size	35	Kier chi6 (path) index
6	inertia moment 1 length	36	Kier chiV3 (path) index
7	inertia moment 2 length	37	Kier chiV4 (path) index
8	inertia moment 3 length	38	Kier chiV5 (path) index
9	ellipsoidal volume	39	Kier chiV6 (path) index
10	total dipole moment	40	Kier chi3 (ring) index
11	dipole moment X component	41	Kier chi4 (ring) index
12	dipole moment Y component	42	Kier chi5 (ring) index
13	dipole moment Z component	43	Kier chi6 (ring) index
14	LogP	44	Kier chiV3 (ring) index
15	total lipole	45	Kier chiV4 (ring) index
16	lipole X component	46	Kier chiV5 (ring) index
17	lipole Y component	47	Kier chiV6 (ring) index
18	lipole Z component	48	Kappa 1 index
19	molecular refractivity	49	Kappa 2 index
20	Kier chi0 (atoms) index	50	Kappa 3 index
21	Kier chiV0 (atoms) index	51	KAlpha 1 index
22	Kier chi1 (bonds) index	52	KAlpha 2 index
23	Kier chiV1 (bonds) index	53	KAlpha 3 index
24	Kier chi2 (path) index	54	shape flexibility index
25	Kier chiV2 (path) index	55	Randić topological index
26	Kier chi3 (cluster) index	56	Balaban topological index
27	Kier chiV3 (cluster) index	57	Wiener topological index
28	Kier chi4 (cluster) index	58	sum of E-state indices
29	Kier chiV4 (cluster) index	5979	2D autocorrelogram (unweighted, separation is from 0 to 20)
30	Kier chi4 (path/cluster) index	80100	2D autocorrelogram (weighting factor is charge, separation is from 0 to 20)

samples which locate nearest position to the gravity are selected from each neuron as a subset.

2.6. Cell-Based Method. A hyperbox including all data points is created in the descriptor space. Each side is divided independently into user specified grid spacing, *h*. The resulting divided subhyperboxes called cell which include any samples are searched, and a sample which locates nearest position to the gravity of the cell is selected as a subset.

EXPERIMENTAL SECTION

Datasets. Dataset I contained 8000 types of tripeptides. Using Biopolymer module of Sybyl V6.3,[30] three-dimensional structures of the tripeptides were constructed. One-hundred (100) chemical and physical descriptors were assigned to each compound by using Tsar V3.1.[28] These descriptors are shown in Table 1. Principle component analysis of the original dataset (8000 × 100) was performed to reduce the redundancy of the data. The first 21 components which explain 95.3% of the variation of the data were selected. The first 21 scores were used as the following analysis.

Dataset II contained 53 064 chemical structures of the Maybridge structural catalog.[31] These two-dimensional chemical structures were converted into three-dimensional structures by using Converter 95.0[32] within Insight II. One-hundred sixteen (116) chemical and physical descriptors were assigned to each compound by using Tsar V3.1. These descriptors are shown in Table 2. Principle component analysis of the original dataset (53 064 × 116) was performed to reduce the redundancy of the data. The first 36 components which explain 95.0% of the variation of the data were selected. The first 36 scores were used as the following analysis.

System Used for Data Analysis. We encoded maximum (dis)similarity method, leader algorithm, k-mean clustering, GAs, Kohonen networks, and cell-based method by FORTRAN programs. All calculations were performed on Indigo 2 running version 6.2 of the IRIX operating system.

RESULTS AND DISCUSSION

Dataset I. Subset Selection. The cell-based method was performed to select diverse subset I. The length of each side of the cell was set to 0.0355. Seven hundred and two (702) of the cells were occupied by the samples. The sample which locate nearest position to the gravity of the cell was selected from each cell as subset I. Maximum similarity method was performed to select redundant subset II with 700 samples. These subsets I and II were selected as the standard of diverse and redundant subsets, respectively.

Maximum dissimilarity method was performed to select subset III with 700 samples. Two types hierarchical clustering methods, group averaging hierarchical clustering method and reciprocal nearest neighbor Ward hierarchical clustering method, were performed to select subsets IV and V. It was impossible to select exactly 700 clusters. So 704 and 701 clusters for group averaging and Ward clustering methods were selected, respectively. The 701 and 704 samples which correspond to the cluster center were selected as subsets IV and V, respectively. Two types k-mean nonhierarchical clustering methods were performed to select subsets VI and VII with 700 samples. When the seed points were selected randomly, the cluster centers were selected as subset VI. When the seed points were selected by using leader algorithm with the threshold being 3.40, the cluster centers were selected as subset VII. GAs were performed to select subset VIII. The average score for each fitness function (MP) was

Table 2. Descriptors in This Study

no.	descriptor	no.	descriptor
1	molecular mass	29	Kier chi6 (path) index
2	molecular volume	30	Kier chiV3 (path) index
3	inertia moment 1 size	31	Kier chiV4 (path) index
4	inertia moment 2 size	32	Kier chiV5 (path) index
5	inertia moment 3 size	33	Kier chiV6 (path) index
6	inertia moment 1 length	34	Kier chi3 (ring) index
7	inertia moment 2 length	35	Kier chi4 (ring) index
8	inertia moment 3 length	36	Kier chi5 (ring) index
9	ellipsoidal volume	37	Kier chi6 (ring) index
10	total dipole moment	38	Kier chiV3 (ring) index
11	dipole moment X component	39	Kier chiV4 (ring) index
12	dipole moment Y component	40	Kier chiV5 (ring) index
13	dipole moment Z component	41	Kier chiV6 (ring) index
14	Kier chi0 (atoms) index	42	Kappa 1 index
15	Kier chiV0 (atoms) index	43	Kappa 2 index
16	Kier chi1 (bonds) index	44	Kappa 3 index
17	Kier chiV1 (bonds) index	45	KAlpha 1 index
18	Kier chi2 (path) index	46	KAlpha 2 index
19	Kier chiV2 (path) index	47	KAlpha 3 index
20	Kier chi3 (cluster) index	48	shape flexibility index
21	Kier chiV3 (cluster) index	49	Randić topological index
22	Kier chi4 (cluster) index	50	Balaban topological index
23	Kier chiV4 (cluster) index	51	Wiener topological index
24	Kier chi4 (path/cluster) index	52	sum of E-state indices
25	Kier chiV4 (path/cluster) index	53−68	2D autocorrelogram (unweighted, separation is from 0 to 15)
26	Kier chi3 (path) index	69−84	2D autocorrelogram (weighting factor is charge, separation is from 0 to 15)
27	Kier chi4 (path) index	85−100	3D autocorrelogram (unweighted, separation is from 0 to 15)
28	Kier chi5 (path) index	101−116	3D autocorrelogram (weighting factor is charge, separation is from 0 to 15)

Table 3. The Results of GAs

initial max.	genrtn mean	min.	final max.	genrtn mean	min.
0.873	0.873	0.872	0.881	0.881	0.881

monitored as the function of generation. The maximum, mean, and minimum scores of the initial (0) and final (50) generations are shown in Table 3. The mean scores of MP increased from 0.873 to 0.881. Each score gradually increased as the number of generations increased. The subset with maximum MP in the final generations was defined as subset VIII. Kohonen networks were performed to select the subset IX. Networks (32 × 32) were constructed, and 725 samples were selected from each neuron as subset IX.

MP Comparison. Prior to box counting analysis, five subsets IV, V, VI, VII, and VIII which were selected by cluster-based methods, were compared from the viewpoint of the clustering efficiency (MP). The MP value of each subset was 0.878, 0.898, 0.877, 0.872, and 0.881 for subsets IV, V, VI, VII, and VIII, respectively. The clustering efficiency MP was decreased as the following order: subsets V, VIII, IV, VI, to VII.

Box Counting Analysis. A hyperbox including all data points was created in the 21-dimensional space. The length of each side was 23.67 ($h = 23.67$). Each side was divided independently by 2 ($h = 11.84$), 3 ($h = 7.89$), 4 ($h = 5.92$), ..., and 400 ($h = 0.0586$). The number of boxes (N_h) containing any data points was counted. The box counting plot of the original dataset I are shown in Figure 2. Linear relations between log h and log N_h were found. Least-squares fitting was performed to estimate D_f for the original dataset. The analyzing range of log h was set between log $h_{min} = -1.23$ and log $h_{max} = 0.675$. D_f of the original dataset was 0.774, while the R^2 of the least-squares fitting was 0.999. Now to compare the data structure of the selected subset and the original dataset, the deviation between log N_h of the original dataset and that of the selected subset (dlog N_h) was estimated. dlog N_h shows the difference in the number of hyperboxes which contain the data of the original dataset without holding the subset on logarithmic scale.

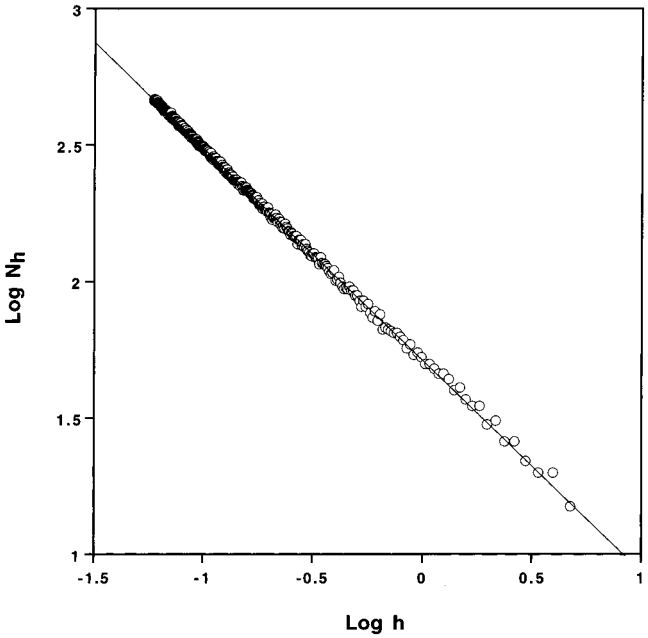

Figure 2. Plot of log N_h vs log h for the original dataset I.

First of all, diverse subset I which was selected by cell-based method and redundant subset II which was selected by maximum similarity method were analyzed. dlog N_h was plotted against log h. The plots of diverse subset I and redundant subset II are shown in Figure 3. dlog N_h of subset I was almost 0 when log h was between 1.5 and −1.0. In this region of log h, the data structure of the original dataset was represented by the subset. This indicates that the subset

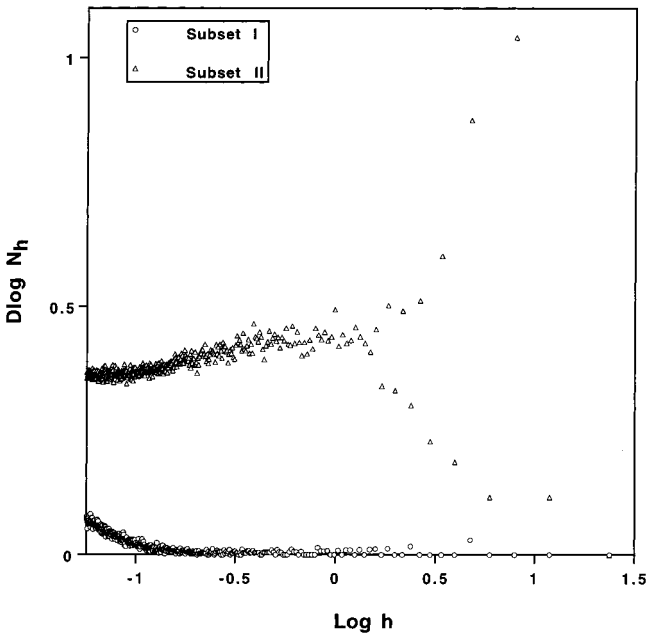

Figure 3. Plot of dlog N_h vs log h for subsets I and II. Subset I was selected by cell-based method. Subset II was selected by maximum similarity method.

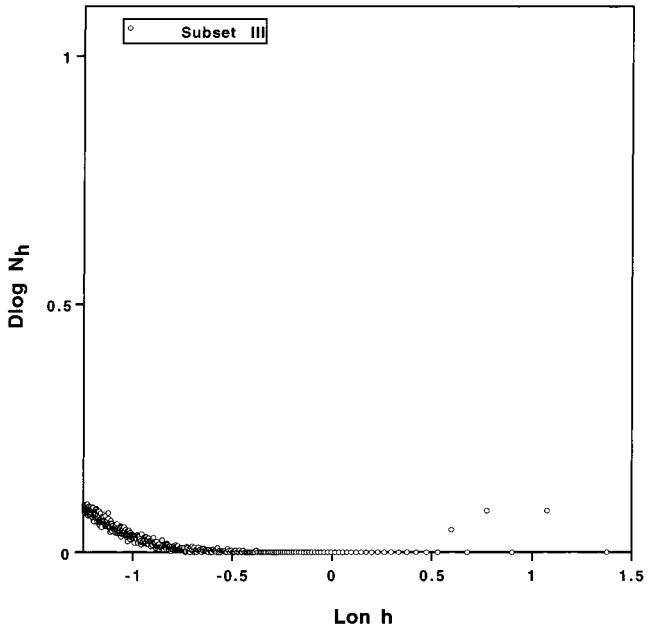

Figure 4. Plot for dlog N_h vs log h for subset III. Subset III was selected by maximum dissimilarity method.

did not lose any information of the original dataset within this region of log h. When log h decreased from -1.0, dlog N_h increased rapidly. The information of the original dataset was lost rapidly in this region of log h. Representativeness of the global data structure of the original dataset was satisfied with subset I. This profile reflects the characteristics of the diverse subset. The profile of redundant subset II was dramatically different from that of subset I. As log h decreased, dlog N_h of subset II showed sharp increase at first. The information of the original dataset was lost within the region of log h. Then dlog N_h decreased. The information of the original dataset was gradually increased within this region of log h. The profile represented the redundancy of subset II. This profile reflects the characteristics of a redundant subset.

The plots of subset III are shown in Figure 4. The profile of subset III resembled that of diverse subset I. dlog N_h of subset III was constantly 0 when log h was between 0.5 and -0.5. In this region of log h, the data structure of the original dataset was perfectly represented by the subset. When log h decreased from -1, dlog N_h increased rapidly. So dataset III is a diverse subset.

The plots of subsets IV, V, VI, VII, VIII, and IX are shown with those of subset I and II in Figures 5−8. The plots of these subsets were located between those of diverse subset I and redundant subset II. The profile of each subset was different. dlog N_h of subsets VI, VIII, and IX increased at first as log h decreased. Then dlog N_h became almost constant. In the first region of log h, the information of the original dataset was lost. In the second region of log h, the data structure of the original dataset was correctly represented by the subsets. Representativeness of the local data structure of the original dataset was satisfied with subsets VI, VIII, and IX. These profiles were different from that of redundant subset II in the following two points. First the increment of dlog N_h of these subsets in the first region of log h was smaller than that of the redundant subset II. Then dlog N_h of these subsets was almost constant or slightly increased but dlog N_h of redundant subset II was decreased in the second region of log h.

dlog N_h of subsets IV, V, and VII increased linearly as log h decreased. This indicated that the information of the original dataset was lost gradually according to decrease log h. These subsets achieved the balance of the representativness of global and local structures of the original dataset. This means these subsets are representative subsets. The profile of subset IV which was selected by group averaging hierarchical clustering method resembled that of subset VII which selected the k-mean nonhierarchical clustering method with the seed points selected by the leader algorithm.

When subsets are selected from the original dataset, some information of the original dataset is retained and some is lost. How information is retained or lost is very important. In this simulation, the cell-based method and the maximum dissimilarity method selected the subset in which data structure reflected global data structures of the original dataset. The subsets were most diverse among all the other selected subsets. On the other hand, k-mean clustering with randomly selected seed points, GAs, and Kohonen networks selected the subsets which data structures reflected local data structure of the original dataset. The data structures of the subsets which were selected by the group averaging clustering method, the reciprocal nearest neighbor Ward clustering method, and the k-mean clustering method with the seed points which were selected by leader algorithm reflected global and local data structure of the original dataset. This means these subsets are representative subsets. The data structures of the subsets selected by the group averaging clustering method and the k-mean clustering method with the seed points which were selected by leader algorithm resembled. These subsets were more diverse than the subset selected by the reciprocal nearest neighbor Ward clustering method.

Dataset II. Subset Selection. The cell-based method was performed to select diverse subset I. The length of each side

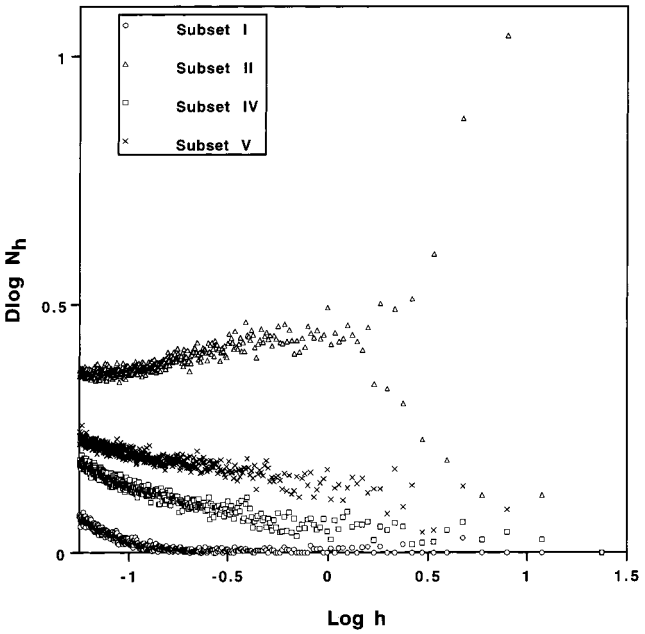

Figure 5. Plot of dlog N_h vs log h for subsets I, II, IV, and V. Subset I was selected by cell-based method. Subset II was selected by maximum similarity method. Subset IV was selected by group averaging clustering method. Subset V was selected by Ward clustering method.

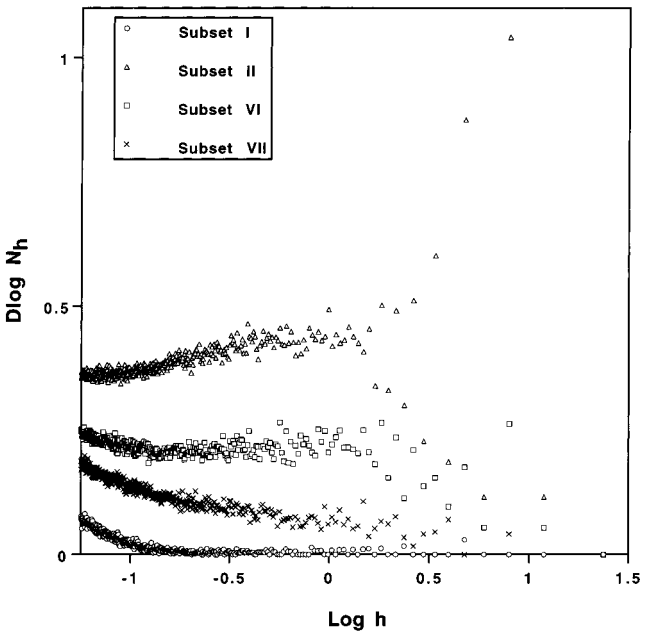

Figure 6. Plot for dlog N_h vs log h for subsets I, II, VI, and VII. Subset I was selected by cell-based method. Subset II was selected by maximum similarity method. Subset VI was selected by k-mean method with the seed points were selected randomly. Subset VII was selected by k-mean method with the seed points were selected by leader algorithm.

of the cell was set to 0.0577. Nine-hundred and ninety-six (996) of the cells were occupied by samples. The sample which located the nearest position to the gravity of the cell was selected from each cell as subset I. The reciprocal nearest neighbor Ward hierarchical clustering method was performed to select subset II with 1003 samples. Two types of k-mean nonhierarchical clustering methods were performed to select subsets III and IV with 1000 samples. When the seed points were selected randomly, the cluster centers were selected as subset III. When the seed points were selected by using leader algorithm with the threshold being 10.21, the cluster centers were selected as subset IV. Kohonen networks were performed to select subset V. Networks (33 × 33) were constructed, and 1036 samples were selected from each neuron as subset V.

MP Comparison. The MP values of the subsets which were selected by cluster-based methods were 0.701, 0.697, and 0.648 for subsets II, III, and IV, respectively. The clustering efficiency MP was decreased in the following order: subsets II, III, to IV.

Box Counting Analysis. A hyperbox including all data points was created in the 36-dimensional space. The length of each side was 150.26 ($h = 150.26$). Each side was divided independently by 2 ($h = 75.13$), 3 ($h = 50.08$), 4 ($h = 37.56$), ..., and 1000 ($h = 0.1502$). The number of boxes (N_h) containing any data points was counted. The box counting plot of the original dataset II is shown in Figure 9. Linear relations between log h and log N_h were found. Least-squares fitting was performed to estimate D_f for the original dataset. The analyzing range of log h was set between log $h_{min} = -0.823$ and log $h_{max} = 1.27$. D_f of the original dataset was 0.730, while the R^2 of the least-squares fitting was 0.999.

dlog N_h was plotted against log h. The plots of the subset I, II, III, and IV are shown in Figures 10 and 11. dlog N_h of subset I was almost 0 when log h was between 2 and −0.8. In this region of log h, the data structure of the original

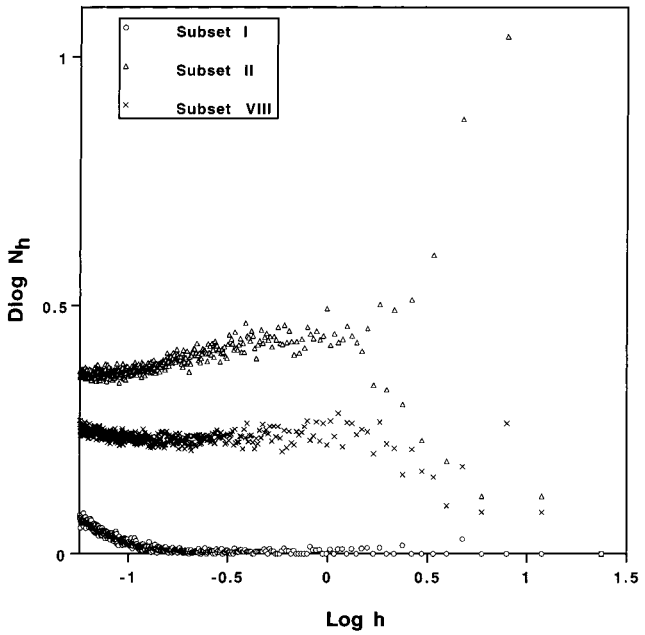

Figure 7. Plot of dlog N_h vs log h for subsets I, II, and VIII. Subset I was selected by cell-based method. Subset II was selected by maximum similarity method. Subset VIII was selected by GAs.

dataset was represented by the subset. When log h decreased from −0.8, dlog N_h increased. The information of the original dataset was lost in this region of log h. Representativeness of the global data structure of the original dataset was satisfied with the subset I. As log h decreased, dlog N_h of subsets II and V increased at first. Then dlog N_h became almost constant. In the first region of log h, the data structure of the original dataset was not represented by the subsets. In the second region of log h, the data structure of the original dataset was correctly represented by the subsets. Representativeness of the local data structure of the original dataset

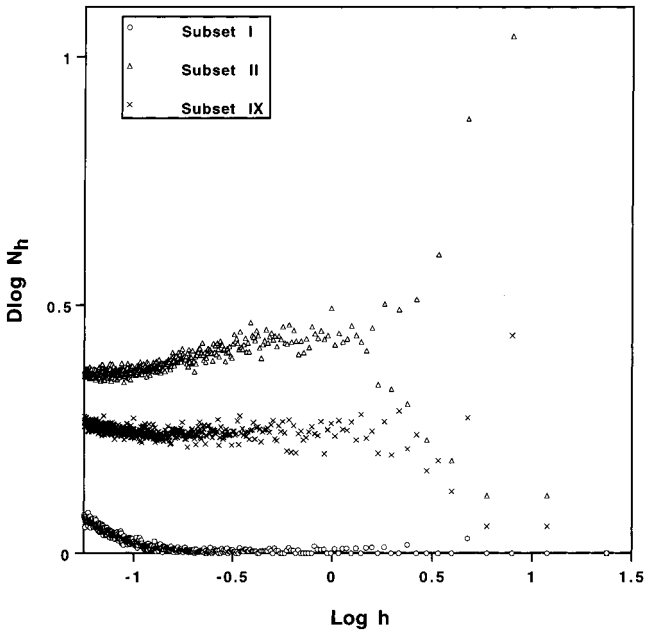

Figure 8. Plot of dlog N_h vs log h for subsets I, II, and IX. Subset I was selected by cell-based method. Subset II was selected by maximum similarity method. Subset IX was selected by Kohonen networks.

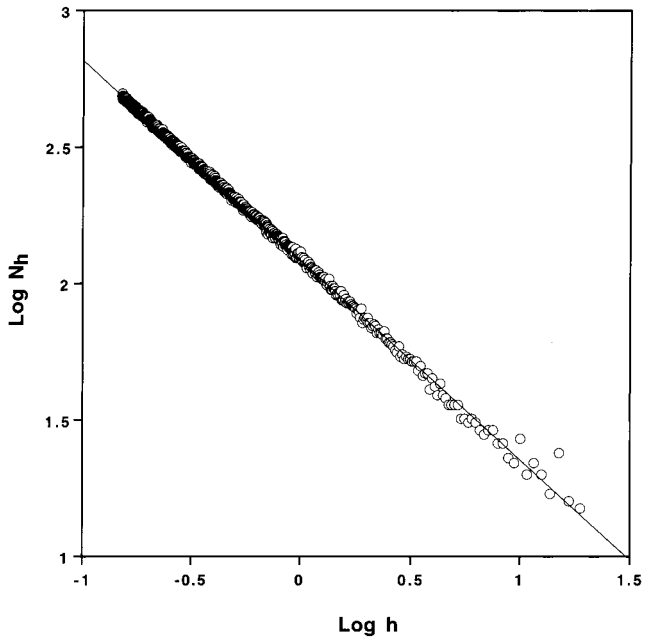

Figure 9. Plot of log N_h vs log h for the original dataset II.

was satisfied with subsets II and V. dlog N_h of subsets III and IV increased linearly as log h decreased. This indicated that the information of the original dataset was lost gradually according to decrease log h.

Through this simulation of relatively large dataset II, the same trends were observed to dataset I. The subset which was selected by the cell-based method was the most diverse of the five selection methods. On the other hand, the k-mean clustering method with randomly selected seed points and Kohonen networks selected the subsets in which data structures reflected local structures of the original dataset. The subsets which were selected by Ward or k-mean clustering method with the seed points were selected by

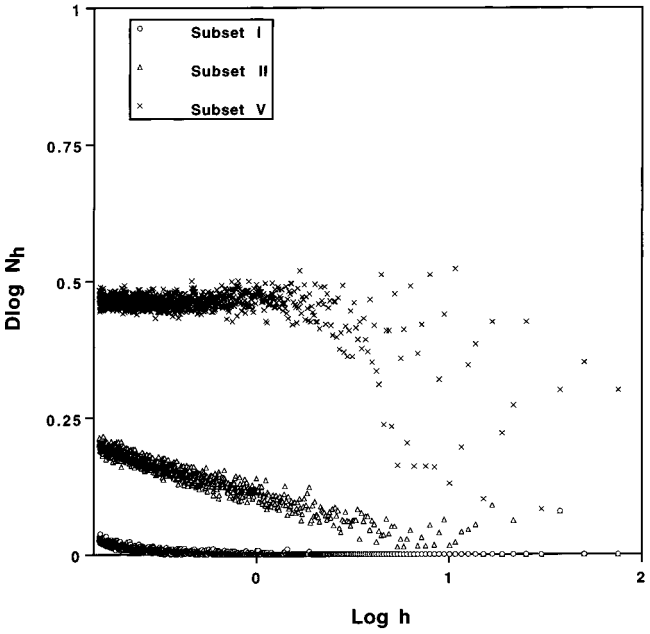

Figure 10. Plot of dlog N_h vs log h for subsets I, II, and V. Subset I was selected by cell-based method. Subset II was selected by Ward clustering method. Subset V was selected by Kohonen networks.

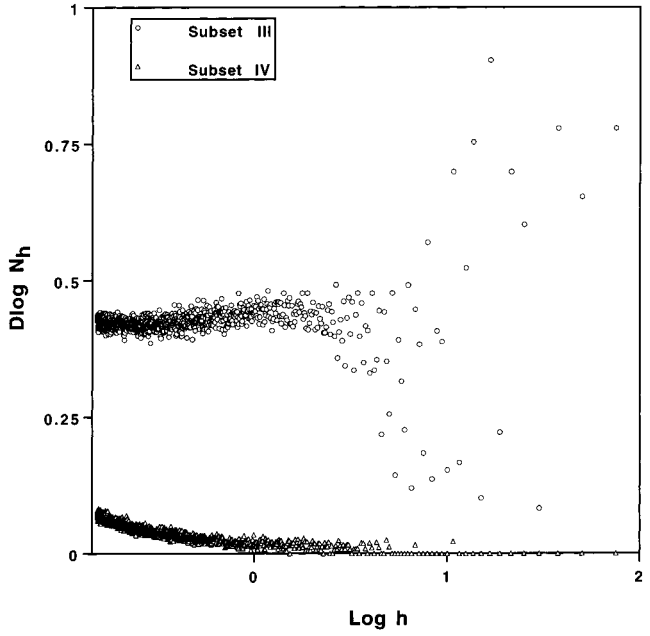

Figure 11. Plot of dlog N_h vs log h for subsets III and IV. Subset III was selected by k-mean method with the seed points were selected randomly. Subset IV was selected by k-mean method with the seed points were selected by leader algorithm.

leader algorithm reflected global and local data structure of the original dataset.

CONCLUSIONS

Complex data structures of the selected subsets from both dataset I with 8000 samples and the dataset with 53 064 were successfully visualized by means of box counting plot. By comparing these plots we could easily understand the characteristics of the data structures. The cell-based method and maximum dissimilarity method selected the subsets which data structure reflected global data structure of the

original dataset. These subsets are diverse subsets. On the other hand, k-mean clustering with randomly selected seed points, GAs with the fitness function consisted of clustering efficiency, and Kohonen networks selected the subsets which data structures reflected local data structure of the original dataset. The data structures of the subsets which were selected by group averaging clustering method, reciprocal nearest neighbor Ward clustering method, and k-mean clustering method with the seed points which were selected by leader algorithm reflected global and local data structure of the original dataset. These subsets are representative subsets. This study demonstrated that the box counting plot is useful to judge how the data structure of a selected subset reflects that of an original dataset. When a new selection method is developed, we could compare the data structure and could understand the characteristics of the selected subsets by using box counting analysis.

REFERENCES AND NOTES

(1) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. M. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(2) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput-Aid. Mol. Des.* **1995**, *9*, 407−416.

(3) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(4) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules from Large Chemical Database. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.

(5) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum Utilization of a Compound Collection or Chemical Library for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892−899.

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

(7) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(8) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topomeric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.

(9) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Deriversed (DPD) Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.

(10) Tominaga, Y.; Fuijwara, I. Data Structure Comparison Using Fractal Analysis. *Chemom. Int. Lab. Syst.* **1998**, *39*, 187−193.

(11) Willett, P. Computational tools for the analysis of molecular diversity. *Perspectives Drug Discovery Design* **1997**, *7/8*, 1−11.

(12) Dunbar, J. B., Jr. Cluster-based selection. *Perspectives Drug Discovery Design* **1997**, *7/8*, 51−63.

(13) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspectives Drug Discovery Design* **1997**, *7/8*, 65−84.

(14) Mason, J. S. Partition-based selection. *Perspectives Drug Discovery Design* **1997**, *7/8*, 1−11.

(15) Mandelbrot, B. The Fractal Geometry of Nature; New York, Freeman, 1983.

(16) Kaye, B. H. A Random Walk through Fractal Dimensions; New York, VCH: 1989.

(17) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing Drug Screening Programs Using Molecular Similarity Methods. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.; New York, 1989; pp 173−176.

(18) Matter, H. Selecting Optimally Diverse Compounds from Structure Database: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(19) Hartigan, J. A. *Clustering Algorithms*; John Wiley & Sons: New York, 1975.

(20) Murtagh, F. A. Survery of Recent Advances in Hierarchical Clustering Algorithms. *Comput. J.* **1983**, *26*, 354−359.

(21) Hibbert, D. B. Genetic Algorithms in Chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(22) Lucasius C. B.; Kateman, G. Understanding And Using Genetic Algorithms Part 1. Concepts, Properties And Context. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(23) Lucasius C. B.; Kateman, G. Understanding And Using Genetic Algorithms Part 2. Representation, Configuration And Hybridization. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.

(24) Kohonen, T. *Self-Organization and Associative Memory*, 2nd ed.; Springer: New York, 1988.

(25) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Newral Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.

(26) Zupan, J.; Gasteiger, J. Neural Networks for Chemists: A Textbook; VCH: Weinheim, 1993.

(27) Barlow, T. W. Self-organizing Maps and Molecular Similarity. *J. Mol. Graph.* **1995**, *13*, 24−27.

(28) Tsar version 3.1, Oxford Molecular Ltd.; Oxford Science Park: Oxford OX4 4GA, UK. (+44) 1865 784600. products@oxmol.co.uk.

(29) Downs, G. M.; Willet, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(30) Sybyl Molecular Modeling Software, version 6.3; Tripos Associates, Inc.: St. Louis, MO 63144.

(31) Maybridge database 1996; Maybridge chemical company Ltd., Trevillett, Tintagel, Cornwall PL34 OHW, UK.

(32) Converter 95.0; MSI 9685 Scranton Road, San Diego, CA 92121-2777.

CI9802070