Molecular Diversity and Representativity in Chemical Databases

Denis M. Bayada,* Hans Hamersma, and Vincent J. van Geerestein

Department of Molecular Design & Informatics, N.V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands

Received June 6, 1998

It is now common practice in the pharmaceutical industry to use molecular diversity selection methods. With the advent of high throughput screening and combinatorial chemistry, compounds must be rationally selected from databases of hundreds of thousands of compounds to be tested for several biological activities. We explore the differences between diversity and representativity. Validation runs were made for different diversity selection methods (such as the MaxMin function), several representativity techniques (selection of compounds closest to centroids of clusters, Kohonen neural networks, nonlinear scaling of descriptor values), and various types of descriptors (topological and 3D fingerprints) including some validated whole-molecule numerical descriptors that were chosen for their correlation with biological activities. We find that only clustering based on fingerprints or on whole-molecule descriptors gives results consistently superior to random selection in extracting a diverse set of activities from a file with potential drug molecules. The results further indicate that clustering selection from fingerprints is biased toward small molecules, a behavior that might partly explain its success over other types of methods. Using numerical descriptors instead of fingerprints removes this bias without penalising performance too much.

INTRODUCTION

High-throughput screening and combinatorial chemistry have radically altered the process of drug discovery. Millions of compounds are offered for testing, and potentially tens of billions of compounds are accessible through combinatorial synthesis. Clearly, even for well-equipped laboratories the latter number is necessarily overwhelming: it is simply impossible, both in available time and allocated resources, to synthesize and test all. Thus, a selection has to be made, either from the commercially available or from the synthetically accessible compounds, which explains the recently increased interest in methods based on the paradigm of molecular diversity. These methods are founded on the assumption, so basic to medicinal chemistry, that similar compounds exhibit similar biological activity;[1] hence, testing a single compound from such a group of similar ones should allow a reasonable estimation of the potential activity of the other cluster members.

For some years now, several research groups have attempted to develop descriptors and selection methods that would facilitate this rational screening process. The reference point is always random selection, and any method or descriptor must be compared not only with other techniques but always with random selection.

In the past, mainly clustering techniques have been used as a selection method and topological fingerprints and pharmacophores as the descriptors.[2] New ways of selections are now emerging that put more emphasis on diversity[3] and less on representativity. Representativity is often understood as selecting compounds that are typical of a particular data set, i.e., they are selected from clusters in the data, one compound for each cluster. Diversity selection is independent of data clusters and because the selection criterion is based on distance to other compounds, it leads predominantly to the selection of outliers. Nowadays, we tend to move away from the fingerprints and pharmacophores and start using whole-molecule numerical descriptors, often those that have been developed previously for QSAR studies.[4] In QSAR, a small set of relatively similar compounds is used, whereas in diversity analyses we typically face the problem of selecting from very large sets of dissimilar compounds. This means that routinely applied QSAR descriptors cannot always be directly applied to diversity selection. As no single descriptor can possibly be sufficient, several different numerical descriptors must be used, and it is important to use only those that have some relevance. Thus, the selection of descriptors must be validated[4−6] by means of a "reference" experiment. As the purpose of diversity selection is to select as small a set of compounds as possible covering as many biological activities as represented by the whole set, descriptors must be linked to biological activity and their performance assessed relative to it.

In this article, we validate some selection algorithms and some molecular descriptors through their ability to discriminate between different biological activities. We will first describe the various descriptors used in this study. Then, selection methods are detailed, and finally the test case and the results are presented.

DESCRIPTORS

Fingerprints. Fingerprints are one of the most popular molecular descriptors. A molecular fingerprint consists of a binary string, where a "one" (on bit) signals the presence of a certain feature from a predefined list (dictionary), whereas a "zero" (off bit) indicates its absence. The implementations vary as to the fingerprints being hashed (more than one feature mapped onto the same bit) or not (one bit corresponding to one feature) and of course to the nature of the features used. We currently use two types of fingerprints. Our implementation of the BCI clustering software[7] uses a

10.1021/ci980109e CCC: $18.00
Published on Web 12/16/1998

Table 1. Descriptors Calculated and Results of Validation for Connection with Biological Activity

name[a]	description[b]	nb[c]	Kept[d]	CSA[e]
clogp	calculated logP	1	yes	20
nbat	number of heavy atoms	1	yes	22
het/nbat	number of heteroatoms divided by the number of heavy atoms	1	yes	21
het	number of heteroatoms	1	yes	22
bonds	number of bonds	1	yes	20
bonds-H	number of bonds between heavy atoms	1	yes	19
molweight	molecular weight	1	yes	20
molw/nbat	molecular weight divided by the number of atoms	1	no	9
rotatable	number of rotatable bonds	1	yes	14
Flex-weight	flexibility index weighted by atomic mass	1	yes	19
Flex-path	flexibility index weighted by bond path	1	yes	19
Surface*	van der Waals surface area	1	yes	20
Surface/nbat*	van der Waals surface area divided by the number of heavy atoms	1	yes	12
Volume*	van der Waals volume	1	yes	21
dB1*, dB2*, ..., dB5*	volx = volume of sphere with radius dx − volume of sphere with radius dx-1 dBx = (number of atoms in volx centered at molecule's centroid)/volx N. B. d0 = 0, d1−5 = 2, 4, 6, 9, 12 Å	5	dB1, dB3, dB4	14, 8, 12, 14, 6
AA1*, AA2*, ..., AA5*	number of acceptor-acceptor pairs at a distance of 2−4, 4−6, 6−9, 9−12, 12−15 Å	5	AA1,2,3,4	19, 19, 20, 17, 9
AD1*, AD2*, ..., AD5*	number of acceptor−donor pairs at a distance of 2−4, 4−6, 6−9, 9−12, 12−15 Å	5	AD1,2,3,4	17, 16, 12, 14, 10
DD1*, DD2*, ..., DD5*	number of donor-donor pairs at a distance of 2−4, 4−6, 6−9, 9−12, 12−15 Å	5	no	8, 10, 10, 9, 2
ratio-don*, ratio-acc*, ratio-neu*	number of donor/acceptor/other atoms by number of atoms	3	no	not used, correlate 0.99 with below
%surf-don*, %surf-acc*, %surf-neu*	percentage of donor/acceptor/other atom surface	3	%surf-acc	8, 17, 9
area-don*, area-acc*, area-neu*	donor/acceptor/other atom surface area	3	yes	15, 22, 20
WHIM*: $\lambda 1$, $\lambda 2$, $\lambda 3$, $\upsilon 1$, $\upsilon 2$, $\kappa 1$, $\kappa 2$, $\kappa 3$, $\gamma 1$, $\gamma 2$, $\gamma 3$	WHIM descriptors, no weighting	11	$\lambda 1$,$\lambda 2$,$\lambda 3$, $\upsilon 1$	14, 18, 18, 13, 10, 4, 5, 9, 6, 4, 4
WHIM-V*: V$\lambda 1$,V$\lambda 2$,V$\lambda 3$,V$\upsilon 1$, V$\upsilon 2$,V$\kappa 1$, V$\kappa 2$, V$\kappa 3$, V$\gamma 1$, V$\gamma 2$, V$\gamma 3$	WHIM descriptors, VDW radius weight	11	V$\lambda 1$, V$\lambda 2$, V$\lambda 3$, V$\upsilon 1$	11, 18, 18, 14, 10, 4, 2, 2, 5, 3, 2
WHIM-M*: M$\lambda 1$, M$\lambda 2$, M$\lambda 3$, M$\upsilon 1$, M$\upsilon 2$, M$\kappa 1$, M$\kappa 2$, M$\kappa 3$, M$\gamma 1$, M$\gamma 2$, M$\gamma 3$	WHIM descriptors, atomic mass weight	11	M$\lambda 2$, M$\lambda 3$, M$\upsilon 1$	10, 19, 19, 12, 10, 2, 2, 4, 5, 3, 2
auto1, auto2, auto3, auto4, auto5, auto6*, auto7*, auto8*, auto9*, auto10*	after computation of the Gasteiger partial charges, an autocorrelation vector is computed for the following 10 distances: 0, 1 bond, 2 bonds, 3 or 4 bonds, 5 or 6 bonds, less than 2 Å, 2−4 Å, 4−6 Å, 6−9 Å, 9−12 Å, 12−15 Å	10	auto1, 2, 3, 4, 5, 6	22, 16, 17, 14, 13, 20, 9, 8, 10, 10

[a] Name of descriptor; an * indicates that the 3D coordinates of the compounds are required to compute this descriptor. [b] Property described. [c] Number of descriptors for each property; [d]Connection of descriptor with biological activity as indicated by CSA. [e] Number of test sets (out of 38) for which CSA indicated the descriptor to be significant.

(user-definable) dictionary of 2465 nonhashed 2D-substructure fragments such as augmented atoms and atom sequences; we do not use frequency information (i.e., setting more than one bit if a certain fragment occurs more than once). In addition, the Chem-X pharmacophore fingerprints[8] are used. These consist of 96 bits storing structural information (including frequency information), plus 10 × 32 bits for the occurrence of various distances, binned into 32 ranges, between the 10 possible pairwise combinations of four pharmacophore types (hydrogen bond donor and acceptor, aromatic ring, and quaternizable nitrogen; additional pharmacophore types, such as lipophilic centers, are presently excluded); these pharmacophore distances are computed for energetically accessible conformations of each molecule and logically combined to produce an overall molecular fingerprint.

Numerical Descriptors. We decided to bias the initial selection of numerical descriptors to those that we considered to be biologically relevant. In all, 86 different descriptors were computed which are listed in Table 1 and described below in some detail.

Descriptor List. Over the years, many different molecular descriptors have been developed. Most of these descriptors apply to QSAR studies of relatively small sets of similar compounds. For current intended use, descriptors should not only describe a meaningful property for hundreds of thousands of potentially diverse compounds but also be fast to compute. Many traditional QSAR descriptors are not applicable for use in diversity as they apply to congeneric series only. As will be shown later in this paper, approximately half of the descriptors we have initially considered were later eliminated.

We started by generating some simple descriptors such as the numbers of atoms and bonds, the molecular weight, and the number of rotatable bonds. Where applicable, descriptors were scaled by dividing by the number of atoms in order to remove correlation with the size of the molecules. As the number of rotatable bonds is not necessarily a good indicator of flexibility, we have developed two new indices,[9] called Flex-weight and Flex-path.

Shape descriptors take into account the three-dimensional shape of the whole molecule. The simplest of these are the surface area and volume. Descriptors dB1 to dB5 count the number of atoms in consecutive spherical layers of specified thickness centered at the centroid of the molecule. There are also distance bin descriptors in which are counted the number of pairs of hydrogen bond acceptors, pairs of one hydrogen bond donor and one acceptor, and pairs of hydrogen bond

donors, the AA, AD, and DD descriptors, respectively. Other descriptors take into account donor, acceptor, and neutral (meaning here neither donor nor acceptor) surface areas as well as the percentages of these surface areas. A donor atom is simply defined as a hydrogen atom linked to a nitrogen or oxygen, and an acceptor atom is defined as a nitrogen of valency smaller than or equal to 3 or an oxygen of valency smaller than or equal to 2.

The WHIM[10] descriptors measure different aspects of the shape of the molecules. In computing these descriptors, the atoms can be unweighted (WHIM), weighted by their van der Waals radii (WHIM-V), or weighted by their atomic weights (WHIM-M). The autocorrelation vector auto[11] takes into account the topology of the molecule (the only exception to our self-imposed rule of not using molecular topology), its three-dimensional conformation, and the Gasteiger partial charges on the atoms.[12]

All descriptors are computed on the basis of the three-dimensional structures of the molecules as obtained by Corina.[13] The logP is computed by the clogp program,[14] the Gasteiger partial charges[15] by the Petra program,[16] and all other descriptors are calculated by in-house programs.

Descriptor Selection. The reason for subset compound selection is to enhance the probability of finding a large spread of biologically active compounds. Therefore, the introduction of descriptors that would describe features unrelated to biological activity would simply generate noise. We have thus attempted to validate the choice of descriptors based on biological activity.

We selected a total of 560 known drugs comprising 38 different biological activities from an in-house database of known drugs, the DrugFile.[17] These 38 activities range from broadly defined activities such as antidepressant and cytostatic, to classes implying a specific mechanism such as oestrogens or 5HT1 antagonists. We also selected a set of 480 compounds from the remaining drugs, i.e., for which none of the 38 selected activity classes were reported in the database.

In all, 1040 compounds have been selected. All descriptors from Table 1 were computed for these compounds. Subsequently, a cluster significance analysis[18] (CSA) was done for each molecular descriptor and each of the 38 sets, i.e., 86 descriptors $\times$ 38 sets $=$ 3268 analyses. The CSA for a given descriptor and a given set S involves comparing the tightness of the cluster of descriptor values of the compounds in S against the tightness of the cluster of values for random sets of size equal to S. The random sets are selected from the whole set, including S. We used the reported[18] threshold value of 0.05 to judge whether a CSA was significant. We then decided that those descriptors that were significant in more than 10 out of 38 test sets were kept (see Table 1). Among the most significant descriptors (significant 18 to 22 out of 38 times) were the number of atoms, the number of heteroatoms, the descriptors involving the acceptor atoms such as the acceptor surface area, the autocorrelation vector, and the $\lambda 2$ and $\lambda 3$ WHIM descriptors (weighted or not). Interestingly, the symmetry (descriptors $\gamma 1,\gamma 2,\gamma 3$) and kurtosis (descriptors $\kappa 1,\kappa 2,\kappa 3$) WHIM descriptors had a significance close to zero, despite being useful in QSAR studies.[10] The descriptors involving donor atoms were not very successful either, but this might be due to our simplistic definition of a donor atom. Eventually, 45 descriptors were selected (see Table 1).

Table 2. Explained Variance and Cumulative Explained Variance for the 10 First Principal Components

principal components	explained variance	cumulative explained variance
1	45.8	45.8
2	12.6	58.4
3	7.4	65.8
4	6.7	72.5
5	3.7	76.2
6	2.9	79.1
7	2.8	81.9
8	2.3	84.2
9	1.9	86.1
10	1.7	87.8

Although other methods are available for assessing the statistical significance of these descriptors,[4−6] we prefer CSA for its independence from linearity in data, its independence from statistical assumptions, and its ease of use. Although, assessing the significance of the descriptors individually is no guarantee that, put together, they will still be meaningful, it is not realistic to try all combinations of 86 descriptor values.

Descriptor Manipulation. The Available Chemicals Directory database[19] (ACD) was selected as being the most diverse database of chemicals at our disposal and thus the one giving the most complete description of chemical property space, of which potential drug molecules are assumed to occupy only a subsection. We, however, decided to delete those molecules that are decidedly undruglike or are very different to drugs we are interested in for our screening program. We eliminated all compounds that had more than 500 atoms (including hydrogen atoms) did not have at least one carbon atom or contained some metals uninteresting to us, such as uranium. Subsequently, a 3D conformer of all remaining compounds was computed using the Corina program. Corina failed to generate a 3D structure for a small number of compounds, which were then also eliminated from further consideration. For the remaining compounds, 45 selected descriptors were computed. At this stage, again a few compounds had to be excluded, mainly because of lack of suitable parameters for clogp and Petra. Finally, 213 430 compounds succeeded this preparative process, which corresponds approximately to 95% of the ACD as of January 1997.

Subsequently, a principal component analysis[20−22] (PCA) study was done using the SAS program.[23] We have arbitrarily selected the first 10 principal components. The average and standard deviation for the 45 descriptors were used for standardization. The first ten PCs explain 87% of the variance, with the first PC explaining 45% and the tenth PC 1.7% (see Table 2).

When we look at the descriptors making up the PCs, we notice that the first three PCs can roughly be understood as describing size, functional groups, and shape, respectively. Other principal components are not easy to interpret.

There are several reasons to use principal component analysis or other similar methods.[24,25] PCA reduces the dimensionality of the data which is very important for graphical inspection of the data. It is, for instance, rather

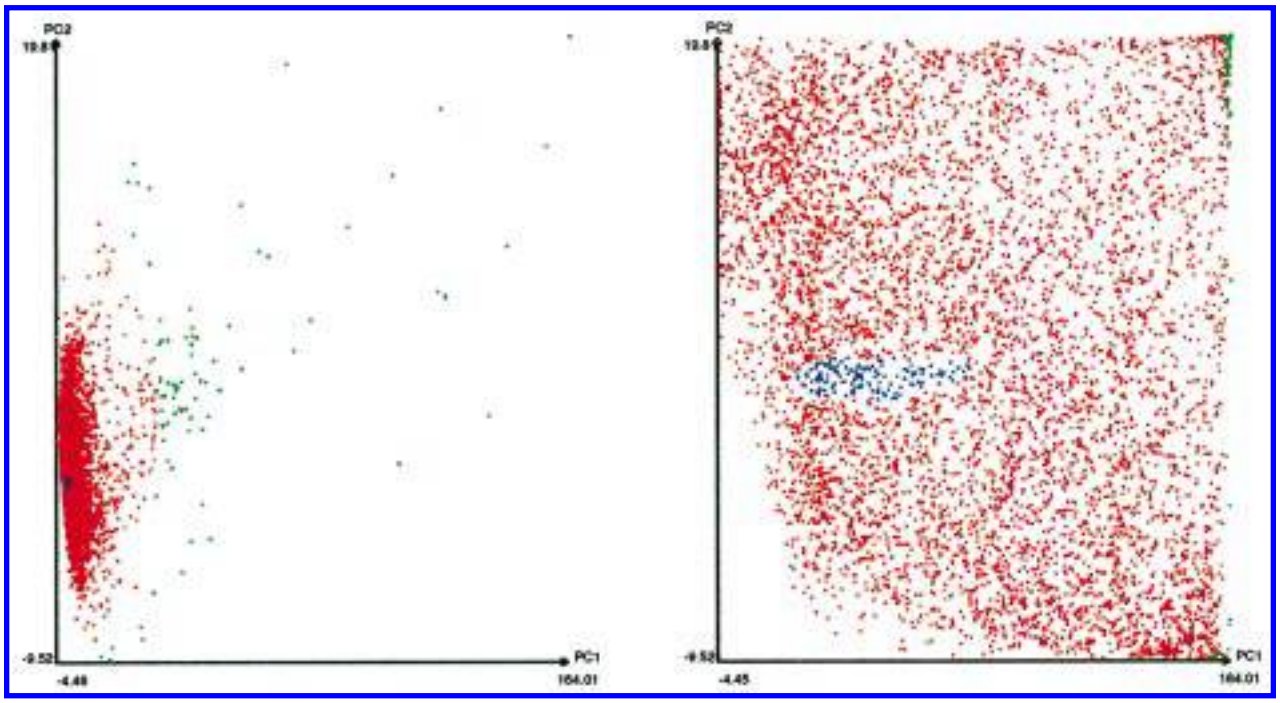

Figure 1. The effect of scaling. Left panel: first two principal component values for our Drug database of 6122 compounds. A set of 81 outliers (green) and a compact set of 134 compounds (blue) have been highlighted. Right panel: the 6122 compounds have been redistributed to produce a uniform density along the two axes (for details, see text). The set of outliers has been swept to the edge of the diagram, whereas the blue set has greatly expanded.

difficult to visualize data in 45 dimensions. Although techniques such as parallel coordinates graphs[26] or Andrews plots[27] allow visualization of data in more than three dimensions, using more than 10 dimensions quickly becomes confusing. However the most important reason for using PCA is the fact that PCA removes the correlations between descriptors. For instance, many of our descriptors correlate heavily with the size of the molecules. This can clearly be seen in that the first PC explains 45% of the variance and mainly takes into account the descriptors linked to molecular size. Traditionally, after a PCA has been applied, the value for each dimension is weighted according to the percentage of explained variance. We have actually chosen not to do so, as we believe this would put too large a weight on size compared to other factors.

CSA was also performed on the PCs. This produced results that were better than for any individual descriptor (results not shown). For that reason, in the rest of this article, the combination of the first ten PCs is used.

With the loadings, averages, and standard deviations, one can normalize any data set uniformly and compute the 10 PCs for this set. This gives a set of 10 values per compound processed. These values are called the PC10 values in the rest of this paper.

Scaling. Normally, the boundaries of descriptor space are determined by outliers, while the bulk of compounds takes up only a small fraction of the large accessible space. For that reason, we devised a method for scaling up the densely populated portions of diversity space and scaling down the sparsely populated diversity space. This is simply done by dividing each dimension in N equally sized bins. In a perfectly sampled space containing M compounds, there should be M/N compounds in each bin. A bin containing P compounds is then resized by a factor of $P/(M/N)$. As a result, dense areas of space expand and sparse areas of space shrink. We have used a value of $N = 100$. When applied to the PC10 values, the scaled values are called scaled_PC10.

As an example of the effect of this nonlinear scaling, Figure 1 is presented. The graph on the left-hand side of Figure 1 is a representation of the first two dimensions of the PC10 values for a set of 6122 drugs used later in this article. Two sets of points are highlighted: in blue, an ensemble of close values in a dense region of space and in green, a set of outliers. After scaling is applied to these data (Figure 1, right-hand side), the outliers, in green, are pushed to the edge of the graph, whereas the small dense region of blue points has clearly expanded. This scaling effect makes the likelihood of selecting compounds in close proximity by a random selection procedure higher than when no scaling was done and vice versa for the outliers. In fact, scaling attempts to give to each compound an equiprobability of being selected by random means. Scaling done in this way makes a diverse selection possible but also a more representative selection is indeed feasible as dense areas of particular interest for drug discovery are well expanded and thus the probability increases that compounds from this area are selected.

This simple type of nonlinear scaling is also a useful aid for simple visualization. In practice, the diversity space is too dense to clearly view data points separately even after applying zooming. Scaling distorts the data in such a way that almost all data points are visible.

METHODS

Clustering. Clustering either assembles entities (agglomerative clustering) or divides sets (divisive clustering) in such a way that the sets produced are as homogeneous as possible; as a result, entities in each subset (cluster) are similar to each other and dissimilar to entities in other clusters. Of the

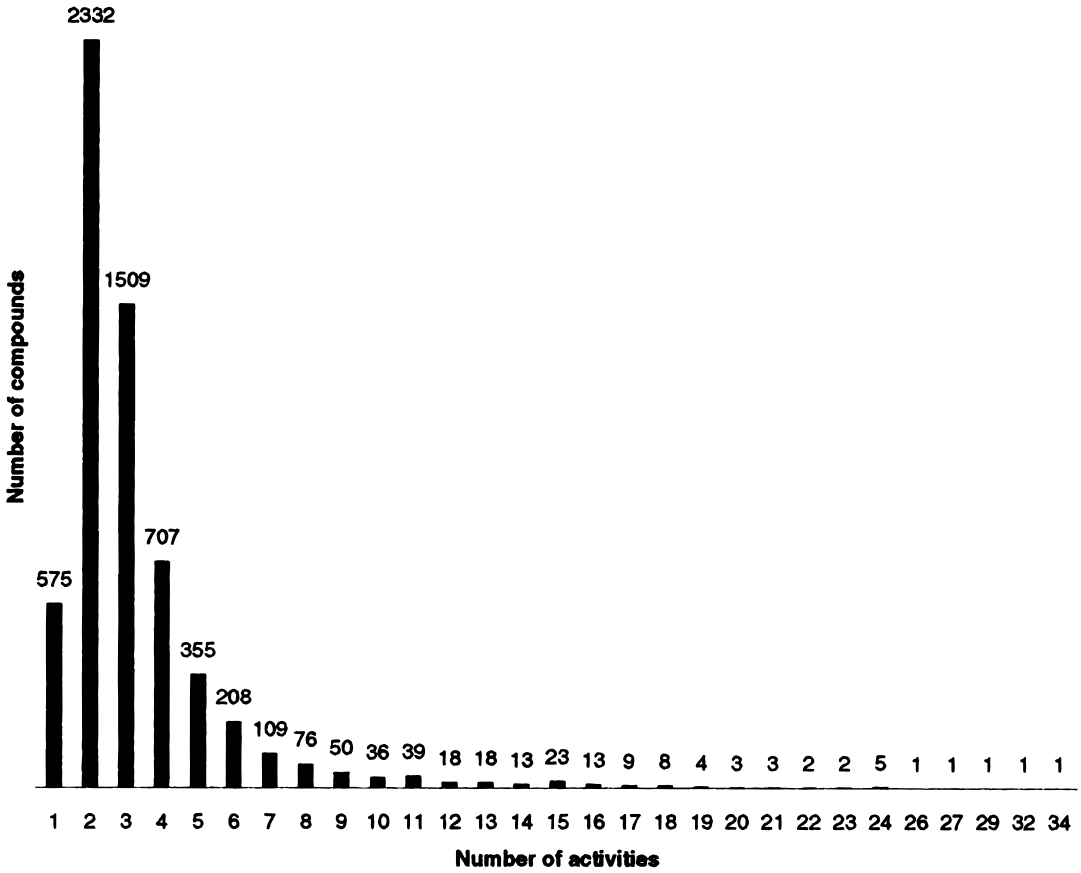

Figure 2. Distribution of activity codes in our DrugFile of 6122 published drugs. The average number of codes is 3.4 activity/compound, while 90% of the molecules have six activities or fewer.

various available algorithms, Ward's method is generally accepted to give the best results.[2] In the BCI implementation of Ward's algorithm, both binary (fingerprints) and numerical data are acceptable as input. The algorithm first creates a list of reciprocal nearest neighbors. This is the time-consuming part, but it needs to be done only once; hereafter, any required number of clusters can be swiftly derived from the reciprocal nearest neighbor list. The cluster members are sorted according to distance from the centroid (the geometrical mean of the cluster members); the molecule closest to the centroid is generally selected as being representative for its cluster.

MaxMin. The MaxMin[3] diversity function maximizes the minimum distance between two selected molecules. This method is dependent on three factors: the number of compounds in the whole set, the number of compounds to select, and the number of dimensions to use. Our implementation of this diversity selection technique consists of repeatedly replacing the two closest selected compounds by two randomly selected compounds up to some point where no improvement can be made after a certain number of cycles. To improve CPU time, we have used the city-block distance $d = \sum|x_i - y_{i-1}|$ rather than the Euclidean distance $d = (\sum(x_i - x_{i-1})^2)^{1/2}$. For a large number of selected compounds (hundreds) in a large set (thousands), we found no significant difference in results obtained with the two distance functions except in terms of CPU time.

Although the selection obtained by MaxMin looks graphically diverse, the CPU time needed for convergence is quite high and precludes its use for very large databases.

Kohonen. A Kohonen neural network[28] is an unsupervised network mapping high-dimensional space into a lower-dimensional space. Densely populated space is mapped onto a larger number of neurons than sparsely populated space. Topological distances are conserved, albeit partially distorted. The method has already been used by Gasteiger's group as a clustering tool for grouping sets of compounds having different activities in nonoverlapping regions of the network.[29]

We used a similar approach for selecting a set of representative compounds. First, a network is constructed with a number of neurons equal to the number of representatives desired. Then the Kohonen neural network is trained using the PC10 values for the data set. After training, each neuron of the network has 10 weights corresponding to the 10 dimensions of the PC10. Subsequently, each compound is put into the neuron whose weights are the closest to its PC10 values, thus spreading the molecules all over the network. Finally, for each nonempty neuron, the compound whose PC10 values are the closest to the neuron's weights is selected, in a similar fashion as the selection of the compound closest to a cluster's centroid. The Kmap program from Gasteiger's group was used to generate the Kohonen maps.

Although this Kohonen selection method is fast (a few minutes at most), the method is not applicable to fingerprints because of their high dimensionality. In a similar fashion as the scaling of data (see Scaling section above), a Kohonen map makes an over-represented area of chemical space easier to target.

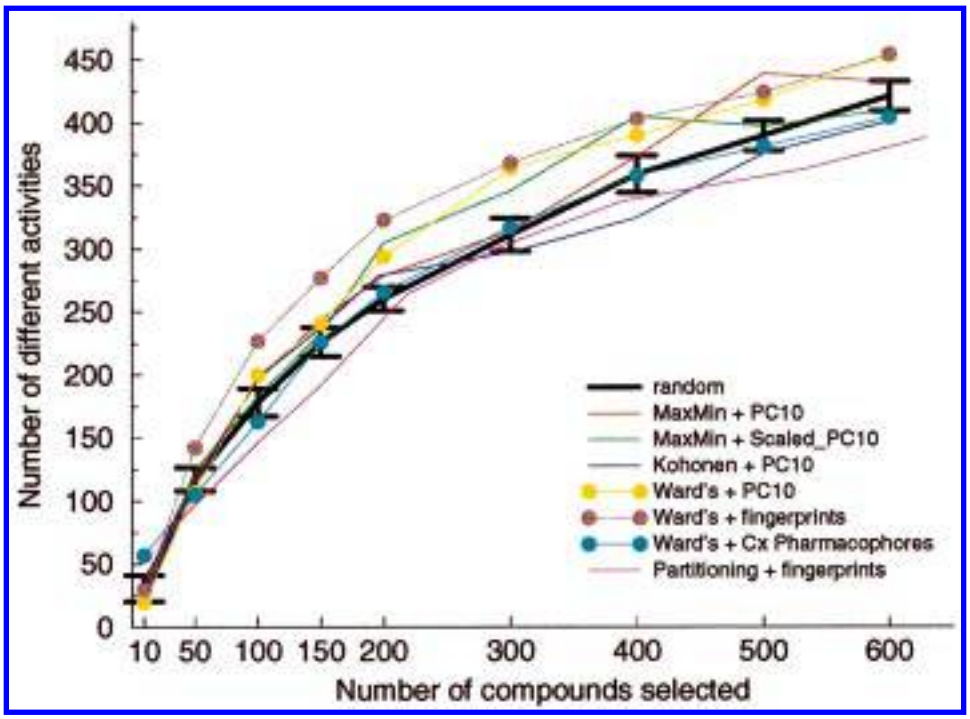

Figure 3. Number of different activity codes present in subsets of various sizes selected from our DrugFile. The subsets were rationally selected using various descriptors and selection methods. The DrugFile contains 6122 published drug molecules with a total of 760 different activities and an average of 3.4 activities per compound.

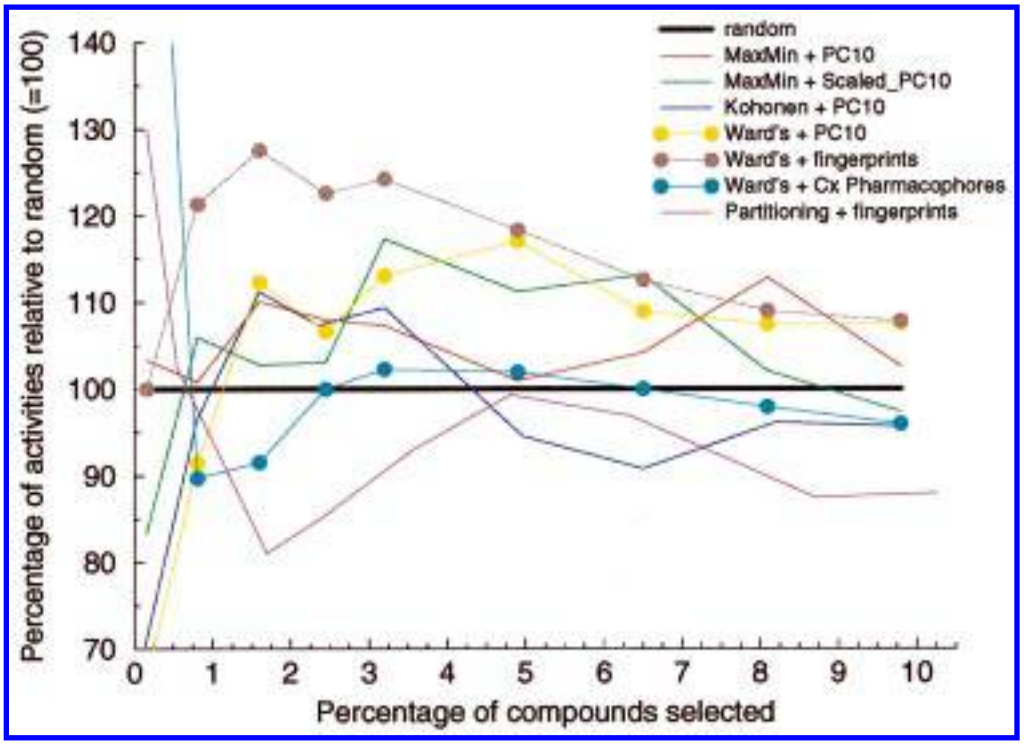

Figure 4. Number of activities extracted from drug file (6122 known drugs, total 760 activities) expressed as a percentage of the number of activities found using random selection. Ward's clustering using 2D-substructure fingerprints or whole-molecule descriptors consistently outperform alternative methods.

Partitioning. We have used a very simple partitioning method similar to the one used by Chem-X for selecting diverse pharmacophores. First, all compounds are put in the list of available compounds. The algorithm randomly selects a first compound and removes this compound as well as all compounds within a specified distance *D* from this compound from the list of available compounds. Then the next compound is randomly chosen from the list of available compounds, and the molecules within a distance *D* from this compound as well as itself are removed from the list. This procedure is repeated until the list of available compounds is empty.

This program must be run several times with different distance thresholds until a convenient number of compounds has been selected. This is no problem as the method is fast and can be run in a few minutes with 20 different threshold

Table 3. Average Number of Activities and Average Number of Atoms for a Selection of 200 Compounds from a File of 6122 Published Drug Molecules Using Different Methods and/or Descriptors[a]

method	descriptors	av no. of activities	av no. of atoms
Ward	fingerprints	4.4	24
MaxMin	PC10	3.1	31
Ward	Chem-x pharm	3.6	24
Ward	PC10	3.3	23
MaxMin	scaled PC10	3.4	25
Kohonen	PC10	3.1	24
partitioning	fingerprints	3.2	28
random		3.3 ± 2.6	25 ± 10

[a] Ward's clustering based on 2D fingerprints finds the highest number of activities, indicating less selective molecules, while the other methods retrieve the average number of activities (average is 3.4 activity/molecule); the MaxMin method applied on the whole-molecule descriptors finds compounds which are significantly larger than those found with other methods.

values on our data set of 6122 compounds (see Results section). For practical reasons, only the BCI fingerprints were used as descriptors for partitioning, although other descriptors such as the PC10 could have also been used.

RESULTS AND DISCUSSION

The following combinations of diversity methods and descriptors were tested:

- MaxMin and PC10
- MaxMin and scaled_PC10
- Kohonen and PC10
- Ward and PC10
- Ward and fingerprints
- Ward and Chem-X pharmacophores
- partitioning and fingerprints

Note that the scaled PC10 is not used with techniques that are dependent on data clusters (Kohonen map, Ward's clustering), simply because clusters disappear after scaling. Highly dimensional descriptors (fingerprints and pharmacophores) are unsuitable for selection techniques such as MaxMin and Kohonen mapping.

Expanding on a previous study,[30] we have used a set of known drugs[17] for which biological activities have been published. This set of 6122 compounds represents 760 different activities with an average of 3.4 activities per compound. Figure 2 shows that most compounds have between one and five activities with a few compounds having more than 10 activities. The goal of this study is to show how effective different selection methods and descriptors can be relative to a random selection in separating biological activities.

As in the previous study,[30] sets with different numbers of compounds have been selected, and for each set, the number of different biological activities were counted. Except for the Kohonen and partitioning techniques, all other methods are capable of selecting any predetermined number of compounds. This was done for 10, 50, 100, 150, 200, 300, 400, 500, and 600 compounds, the latter representing approximately 10% of the data set. For Kohonen and partitioning, first a series of selections was made. We then used those selections where the number of compounds came closest to the size of the subsets from the other methods.

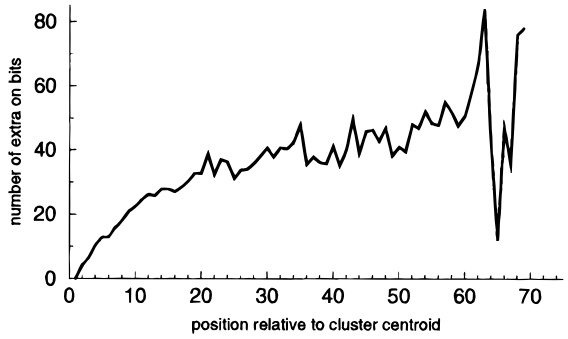

Figure 5. Average number of extra bits per compound vs the rank number relative to compound closest to centroid. Data based on 6122 compounds divided over 200 clusters with an average of 30.6 compounds per cluster; the average number of on bits per compound is 123 ± 41, the average for the cluster centroids = 112 ± 41.

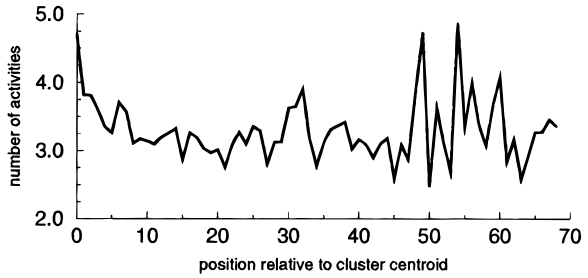

Figure 6. Average number of activities per compound vs the rank number relative to compound closest to centroid (0 = closest to centroid, 1 = next closest, etc.) Data based on 6122 compounds divided over 200 clusters with an average of 3.3 ± 2.6 activities per compound. As there are only 19 clusters with more then 50 compounds data above 50 give an indication only. The average number of activities for the most central compound is 4.7 ± 4.5.

The methods we have applied can be grouped into two diverse selection methodologies (MaxMin+PC10 and Partitioning+fingerprints) and five representative selection methodologies. Clustering is always based on representativity, while Kohonen maps and data scaling give more weight to more populated regions of space and thus can be categorized as representative selection.

From Figures 3 and 4, it is evident that Ward's+fingerprints does better than any other method, with Ward's+PC10 coming a close second. Although little improvement relative to random selection happens after the 300 compounds mark (5% of data set), it must be noticed that to select up to approximately 400 activities, random selection needs almost 50% more compounds than Ward's+fingerprints.

Techniques other than Ward's+fingerprints and Ward's+PC10 that do quite well are MaxMin+PC10 and MaxMin+Scaled_PC10. Nevertheless, these behave very irregularly because of the intrinsic randomness in our implementation of MaxMin. What is striking in Figures 3 and 4 is the apparent failure of the other methods (Kohonen, Pharmacophores, and Partitioning), which sometimes do worse than random.

It would seem from inspecting Table 3 that using representative techniques results in selecting smaller compounds than when using diversity approaches. This can be understood by the fact that diversity methods have the tendency of selecting outliers, which often are relatively large compounds. The ability of the fingerprints in combination with Ward's clustering for selecting compounds with a large

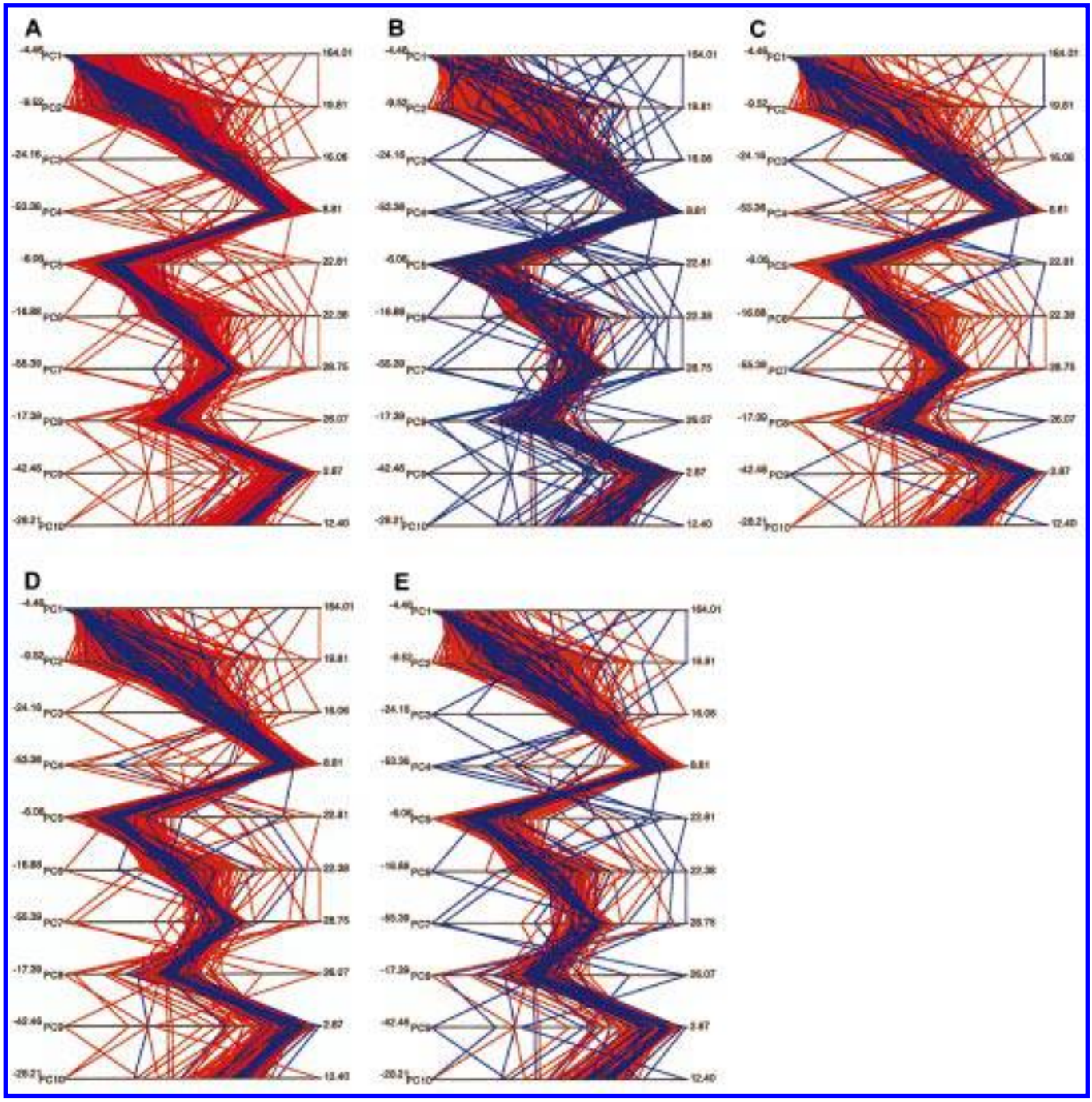

Figure 7. Parallel coordinates representation of first 10 principal components of whole-molecule descriptors for 6122 drug molecules (red lines). Blue lines show subsets of 50 compounds selected using various techniques: A, Random; B, MaxMin + PC10; C, MaxMin + Scaled PC10; D, Kohonen + PC10; and E, Ward's + PC10. See text for further discussion.

number of different activities in comparison to the other combinations of techniques and descriptors is surprisingly good. The compounds not only have more activities but also are smaller in number of atoms. This is due to selecting those compounds that are the closest to the centroid. As is illustrated in Figure 5, the closer a compound is to the centroid, the smaller is the number of on-bits in the fingerprints. As there is some correlation (at least for similar compounds) between the number of on-bits in the fingerprints and the number of atoms, selecting compounds close to centroids is in effect similar to selecting molecules with relatively small number of atoms.

This can be demonstrated in the following example. A cluster of three molecules is composed of molecule A (binary fingerprint 1, 0, 1, 0), molecule B (1, 1, 1, 0) and molecule C (1, 0, 1, 1). Clearly there is one bit difference between A and B and A and C and two bits difference between B and C. The centroid of this cluster is 1, 0.33, 1, 0.33, and compound A is the closest to the centroid. If more compounds are added to this cluster, they will probably have bits 1 and 3 on and some other bits too. Compound A with bits 1 and 3 on is a core compound from which one can make the other molecules just by adding functional groups that will turn some other fingerprint bits on. In the more general case, a small core compound is closer to the centroid because it lacks features that other molecules in the same cluster but not the majority of them have.

This is an important side-effect of using clustering that should not be ignored. In general, smaller compounds tend to be less selective and thus should have more biological activity codes. This is shown in Figure 6, which illustrates that the average number of activities for a compound decreases with increased distance to the compound the closest to the centroid of a cluster. In Figure 6, the curve is not significant for large clusters because the number of these clusters is small. If the goal of selection through Ward's+ fingerprints is to screen the selected compounds for potential biological activities, more hits might be found than by using other methods, but the overall quality of the hits will be lower, i.e., nonselective compounds will be found.

Another explanation for the preference of clustering for selecting small compounds might be a bias in our data set. Obviously, all data sets are in some way biased as no absolute chemically random set of compounds exists. Chemicals and especially drugs databases are made by chemists that work in some non-random fashion. Once a hit is known, chemists tend to add and remove functional groups from this hit to explore the chemical/biological universe around it. This means that many compounds are usually known around a well-defined hit. This indeed introduces a strong bias in databases where clusters of similar molecules have been introduced non-randomly. Thus one compound from each of these artificial clusters would be selected by a clustering program used in combination with topological fingerprints. This might partially explain the success of such a technique.

The choice of selecting compounds through diversity or representativity techniques is not so simple. In fact there is not just one type of representativity. Clustering can be seen as minimal representativity where only one compound per cluster of similar compounds is chosen. Scaling data and Kohonen neural network give maximal representativity in that outliers are usually completely discarded and several compounds can be selected from very dense clusters. This behavior can be illustrated using parallel coordinates[26] illustrations of the data set under study (Figure 7). In a parallel coordinates representation, the coordinate axes are shown in a user-defined order, parallel rather than orthogonal to each other. An n-dimensional point in orthogonal coordinates is shown in parallel coordinates as a broken line composed of $n-1$ segments. A point P whose coordinates are (P1, P2, ..., Pn) is thus represented by $n-1$ segments (Si) whose coordinates are (Pi,P$i+1$) for $i = 1$ to $n-1$. The blue lines represent selections of 50 compounds from the whole data set (in red) by different methods. The methods involving fingerprints or pharmacophores are not shown. Similarly to random selection (Figure 7A), MaxMin+Scaled (Figure 7C) and Kohonen (Figure 7D) selected almost exclusively from the most densely populated region of space. MaxMin+PC10 (Figure 7B) selects almost all outliers and clearly seems to be the most diverse selection. Ward's clustering+PC10 (Figure 7E) is a compromise between selection of outliers and selection from the densely populated regions of space.

Our preferred technique is Ward's clustering on the PC10 values for diversity analysis of large and diverse databases. Ward's clustering has the advantage over many other selection methods of being quite fast, though nowhere as fast as Kohonen networks or partitioning, and of being a good compromise between a purely diverse and an over-representative selection. The PC10 values have also a few advantages over binary fingerprints. They can be easily visualized using traditional two- or three-dimensional graphs or parallel coordinates systems. Being numerical, they can be manipulated with ease, and missing values can readily be generated. They are almost always valid, even for odd compounds, contrary to the fingerprints that might not encode a given molecule's rare chemical functionality. This is because fingerprints are based on a predefined dictionary, which in turn is based on fragments generated from known molecules. New molecules outside this known set may contain new fragments which are not represented in the dictionary; hence no bit can be set in the fingerprints, and these features are thus not taken into account in diversity selection. Finally, the combination of Ward's clustering and PC10 values appears to avoid the selection of small nonselective compounds.

CONCLUSIONS

We have explored different selection techniques and molecular descriptors that allow diverse or representative selection of subsets of compounds from a database. As has been noted previously, we have also found that Ward's clustering in association with topological fingerprints appears to be the most effective method to select a set of biologically active compounds from a large compound collection. However, this type of selection seems to be biased toward selecting small molecules. Other techniques such as MaxMin, Kohonen networks, and partitioning failed to give good results, as did scaled descriptors and pharmacophore fingerprints. By contrast, using Ward's clustering with principal components derived from validated whole-molecule numerical descriptors results close to those using topological fingerprints are found without a bias toward small molecules.

ACKNOWLEDGMENT

The authors wish to thank Christof Schwab for modifying the Kohonen mapping program, Kmap, and providing judicious advice on its use.

REFERENCES AND NOTES

(1) Maggiora, G. M.; Johnson, M. A. Concepts and Applications of Molecular Similarity; John Wiley & Sons: New York, 1990.
(2) 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.
(3) Hassan, M.; Bielawaski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Diversity* **1996**, *2*, 64−74.
(4) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors *J. Med. Chem.* **1997**, *40*, 1219−1229.
(5) 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.
(6) Delaney, J. S. Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Diversity* **1995**, *1*, 217−222.
(7) Barnard, J. M.; Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(8) Chem-X user manual; Oxford Molecular Ltd, The Medawar Centre, The Oxford Science Park, Oxford OX4 4GA, U.K.
(9) Knegtel, R. M. A.; Bayada, D. M.; Engh, R. A.; von der Saal, W.; van Geerestein, V. J.; Grootenhuis, P. D. J. Comparison of two implementations of the incremental construction algorithm in flexible

docking of thrombin inhibitors. *J. Comput.-Aided Mol. Design*. Accepted for publication.

(10) (a) Todeschini, R.; Lasagni, R.; Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemometrics*, **1994**, *8*, 263−272. (b) Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons., Chemometrics and intelligent laboratory systems. *Chemometrics and Intelligent Laboratory Systems* **1995**, *27*, 221−229. (c) Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Quantum Struct.-Act. Relat.* **1997**, *16*, 113−119.

(11) (a) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: Perception, autocorrelation descriptor and SAR studies: Autocorrelation descriptor. *European J. Med. Chem.* **1984**, *19*, 66−70. (b) Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: Perception, autocorrelation descriptor and SAR studies: Use of the autocorrelation descriptor in the qsar study of two nonnarcotic analgesic series. *European J. Med. Chem.* **1984**, *19*, 79−84.

(12) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.

(13) Sadowski, J.; Wagener, M.; Gasteiger, J. CORINA: Automatic Generation of High-Quality 3D-Molecular Models for Application in QSAR. In *10th European Symposium on Structure−Activity Relationships: QSAR and Molecular Modelling*; Sanz, F., Ed.; Prous Science Publishers: 1994.

(14) BioByte, 645 N. College Ave., Claremont, CA 91711, U.S.A.

(15) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, 1988; pp 119−138.

(16) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges *Tetrahedron* **1980**, *36*, 3219−3228.

(17) Primary sources are Drugs of the Future, Drug Data Report, MartinDale, Pharmaprojects, USP Dictionary of USAN and International Drug Names.

(18) McFarland, J. W.; Gans, D. J. On the significance of clusters in the graphical display of structure−activity data. *J. Med. Chem.* **1986**, *29*, 505−514.

(19) Molecular Design Limited, 14600 Catalina Street, San Leandro, CA 94577.

(20) Martin, E. J.; Blaney, E. J.; 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.

(21) Gibson, S.; McGuire, R.; Rees, D. C. Principal components describing biological activities and molecular diversity of heterocyclic aromatic ring fragments. *J. Med. Chem.* **1996**, *39*, 4065−4072.

(22) Massart, D. L.; Vandeginste, B. M. G.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. Handbook of chemometrics and qualimetrics: Part A; Elsevier Science B. V.: Amsterdam, 1997; pp 519−556.

(23) SAS 6.10, SAS Institute Inc., Cary, NC.

(24) Cummins, D. J.; Websters Andrews, C.; Bentley, J. A.; Cory, M. Molecular diversity in chemical databases: Comparison of medicinal chemistry knowledge bases and databses of commercially avalaible compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(25) Domine, D.; Devillers, J.; Chastrette, M. A Nonlinear map of substituent constants for selecting test series and deriving structure−activity relationships. 1. Aromatic series. *J. Med. Chem.* **1994**, *37*, 973−980.

(26) (a) Wegman, E. J. Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc.* **1990**, *85*, 664−675. (b) Inselberg, A.; Dimsdale, B. Multidimensional lines I: Representation. *SIAM J. Appl. Mathematics* **1994**, *54*, 559−577.

(27) Andrews, D. F. Plots of high-dimensional data. *Biometrics* **1972**, *28*, 125−136.

(28) Li, X.; Gasteiger, J.; Zupan, J. On the Topology Distortion in Self-Organizing Feature Maps. *Biol. Cybern.* **1993**, *70*, 189−198.

(29) Wagener, M.; Sadowski, J.; Gasteiger, J. Assessing Combinatorial Libraries with Spatial Autocorrelation Functions and Neural Networks. In *Software-Entwicklung in der Chemie 10*; Gasteiger, J., Ed.; GDCh, Frankfurt am Main, 1996; pp 371−380.

(30) Geerestein, V. J. van; Hamersma, H.; Helden, S. P. van Exploiting molecular diversity: pharmacophore searching and compound clustering. in: Computer-assisted lead finding and optimization. In *Current tools for medicinal chemistry*; Waterbeemd, H. van de, Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, 1997; pp 159−178.

CI980109E