Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection

Robert D. Brown*,† and Yvonne C. Martin‡

Pharmaceutical Products Division, Abbott Laboratories, D47E/AP10, 100 Abbott Park Road, Abbott Park, Illinois 60064-3500

Received September 6, 1995⊗

An evaluation of a variety of structure-based clustering methods for use in compound selection is presented. The use of MACCS, Unity and Daylight 2D descriptors; Unity 3D rigid and flexible descriptors and two in-house 3D descriptors based on potential pharmacophore points, are considered. The use of Ward's and group-average hierarchical agglomerative, Guénoche hierarchical divisive, and Jarvis−Patrick nonhierarchical clustering methods are compared. The results suggest that 2D descriptors and hierarchical clustering methods are best at separating biologically active molecules from inactives, a prerequisite for a good compound selection method. In particular, the combination of MACCS descriptors and Ward's clustering was optimal.

INTRODUCTION

The advent of high-throughput biological screening methods have given pharmaceutical companies the ability to screen many thousands of compounds in a short time. However, there are many hundreds of thousands of compounds available both in-house and from commercial vendors. Whilst it may be feasible to screen many or all of the compounds available, this is undesirable for reasons of cost and time and may be unnecessary if it results in the production of some redundant information. Therefore there has been a great deal of interest in the use of compound clustering techniques to aid in the selection of a representative subset of all the compounds available.[1−8] A similar problem faces those interested in designing compounds for synthesis; a good design will capture all the required information in the minimum number of compounds.[9]

Underpinning the compound selection methods is the *similar property principle*[10] which states that structurally similar molecules will exhibit similar physicochemical and biological properties. Given a clustering method that can group structurally similar compounds together, application of this principle implies that the selection, or synthesis, and testing of representatives from each cluster produced from a set of compounds should be sufficient to understand the structure−activity relationships of the whole set, without the need to test them all.

An appropriate clustering method will, ideally, cluster all similar compounds together whilst separating active and inactive compounds into different sets of clusters. The first factor will ensure that every class of active compound is represented in the selected subset but that there is no redundancy. The second factor will minimize the risk that an inactive compound is selected as the representative of a cluster containing one or more actives, thereby missing a class of active compounds.

Clustering is the process of dividing a set of entities into subsets in which the members of each subset are similar to each other but different from members of other subsets. There have been numerous cluster methods described; general discussions of many of these are given by Gordon,[11] by Everett,[12] and by Sneath and Sokal.[13] Several of these methods have be applied to clustering chemical structures; comprehensive reviews are given by Barnard and Downs[14] and by Downs and Willett.[15] In outline, the clustering process for chemical structures is as follows.

(1) Select a set of attributes on which to base the comparison of the structures. These may be structural features and/or physicochemical properties.

(2) Characterize every structure in the dataset in terms of the attributes selected in step one.

(3) Calculate a coefficient of similarity, dissimilarity, or distance between every pair of structures in the dataset, based on their attributes.

(4) Use a clustering method to group together similar structures based on the coefficients calculated in step three. Some clustering methods may require the calculation of similarity values between the new objects formed and the existing objects.

(5) Analyze the resultant clusters or classification hierarchy to determine which of the possible sets of clusters should be chosen.

A number of methods are available both for the production of descriptors in steps (1) and (2) and clusters in step (4). Whilst there are also a large number of coefficients that might be used in step (3), the choice of clustering method may determine which is best suited.

In this paper we present a study aimed at identifying the most suitable descriptors and clustering methods for use in compound selection. We have used a variety of methods to cluster sets of structures with known biological activities and evaluated the clusters produced according to their ability to separate active and inactive compounds into different sets of clusters. We have concerned ourselves with structure based clustering. For this, the *substructure search screens* used in commercial database searching software have often been used as descriptors. We have examined a number of these descriptors, together with two developed in-house, and have considered the use of four commercially available clustering methods.

† brownr@abbott.com.
‡ yvonne.martin@abbott.com.
⊗ Abstract published in *Advance ACS Abstracts,* January 15, 1996.

0095-2338/96/1636-0572$12.00/0

STRUCTURAL DESCRIPTORS

The descriptors examined in this study can be divided into two types: those that characterize the two-dimensional structure of a molecule and those that consider its three-dimensional structure.

2-D Screens. 2-D screens may again be divided in two classes: *structural keys* and *fingerprints*.

Structural Keys. Structural keys were developed for the first substructure search systems. The key for a structure is an array of Boolean values, each of which represents the presence or absence of a specific 2D fragment. Structural keys rely on the use of a predefined *fragment dictionary*, which is a list of fragments selected to be of importance. A series of studies established the types of fragments that should be indexed and that for maximum screening efficiency the set of fragments selected should obey two principles.[16−18]

(1) Independence of occurrence: If two fragments do not occur independently, redundant information is included reducing the efficiency of the screen.

(2) The fragments should be of approximately equifrequent distribution. Frequently occurring fragments will not be useful in distinguishing structures, and very infrequently occurring fragments are unlikely to occur in many queries and so will be rarely used. An unequal distribution of frequencies will lead to poor screenout.

In the case of substructure searching a lack of independence merely introduces redundancy. When using these screens for similarity calculations a lack of independence is more serious. Calculating similarities on the basis of co-occurring fragments will bias the coefficient in favor of the feature in common in those fragments since they will have an increased weight. Unequal frequencies are less of a problem for similarity calculations. When considering their effectiveness as descriptors for similarity calculations, it is important to bear in mind that these fragment sets are selected by criteria aimed at maximizing substructure search screenout.

Within any large set of structures the number of fragments is huge. The fragment sets tend to have a highly skewed distribution with a few fragments occurring in almost all compounds and many occurring only in a few. Furthermore, many fragments co-occur since they describe many of the same atoms and bonds. Dictionary building has therefore generally been an intellectual, manual process in which this fragment set is reduced by selecting fragments on the basis of the rules described above.

The structural key descriptors used in our experiments were taken from the MACCS substructure search system.[19] This uses 960 keys, of which 166 may be used for *key searches* in MACCS and are hence defined in the manual.[20] We examined the use of both the full 960 keys and the smaller set. Characterizations of structures using the full set of keys were obtained using an ISIS application known as the SSKEYS gateway. This returns a 960-bit string for each input structure, indicating the presence or absence of each key.

We used 153 keys of the published 166 to generate a second structural key descriptor. These keys consist of a range of types of generic and specific fragments. The Unity DBMKFPRINT program[21] was used to generate the descriptors, and the 153 were those that could be easily represented in Tripos' SLN notation.[22] With structural keys it is possible to record not only the presence or absence of each fragment but also the number of occurrences. Since it was necessary to use a binary encoding for input to the clustering programs, a number of bits in a bit-string were assigned to each key, allowing the number of occurrences of each fragment up to that number to be recorded. The use of 1 occurrence (153-bit string), 3 occurrences (459-bit string), and 13 occurrences (1989-bit string) was examined. The Daylight fingerprint format which was used throughout these experiments, imposes an upper limit of 2048 bits, hence 13 occurrences was the maximum that could be recorded. In this paper we will refer to this subset as the *MACCS Keys* and the larger set as the *SSKEYS*.

Hashed Fingerprints. Hashed fingerprints were designed to address a fundamental disadvantage of structural keys, their lack of generality. If any database structure contains few of the dictionary fragments, then it will be poorly characterized and hence unlikely to be screened out in a substructure search. Fingerprints dispense with the fragment dictionary and instead define a set of patterns to index. An exhaustive list of all fragments in a structure matching that pattern is then generated and all fragments encoded in a bit-string. A pattern might, for example, be a path of length seven bonds. A path of length n bonds is defined as $(\text{atom-bond-atom})_n$ where the nature of the atoms and bonds at each point is recorded.[23]

The number of distinct fragments produced from any structure of nontrivial size is obviously very large, and the fragment set differs from molecule to molecule. Therefore it is not possible to assign a separate bit in a string to each fragment. Instead, each fragment is *hashed* (that is, used as the seed for a pseudo-randomizing algorithm) to produce a pattern of a few bits over a bit-string of a predefined length. Thus, a given path will always set the same pattern of bits. However, collisions can and will occur when two different paths cause some of the same bits to be set. Whilst this reduces the accuracy of the fingerprints stored compared to a structural key, the much greater number of fragments encoded ensures that, for substructure searching at least, the overall efficiency of the characterization is increased. The design of the algorithm is such that the chances of two different patterns setting the same set of bits is extremely low.[23]

For use in similarity calculations, fingerprints have the conceptual problem that any individual bit does not actually represent anything unique about the structure of the molecule. Furthermore, a similarity coefficient that compares the bits set in two strings may see the same bit position as set and so record a contribution to the similarity coefficient, when in fact that position was set by a different fragment in each of the two structures. Whilst the chance that two different patterns produce the same set of bits is very low, the chance of many different patterns setting the same bit is much higher.

We examined two hashed fingerprints in our experiments, those from the Daylight[24] and Unity[21] systems. The Daylight fingerprints encode each atom's type, all augmented atoms and all paths of length 2−7 atoms. An augmented atom encodes a central atom and the nature of the atoms and bonds attached to it. An option exists to *fold* fingerprints. We could not use this option due to constraints imposed by some of the clustering algorithms examined so instead we used a fix length 1024-bit string.

Whilst the Daylight fingerprinting algorithm hashes all recorded information over the whole length of the fingerprint,

the method used in Unity is to keep information from different length paths distinct. Thus, separate regions of a bit string record information from paths of lengths 2, 3, and 4 including hydrogens and 4−6 excluding hydrogens. A few generic structural keys are added for some common atom and ring systems counts, producing a total string of 988 bits.

3D Screens. Several 3D substructure search systems use a screening stage prior to geometric search.[21,25,26] The screening methods used in these systems encode the spatial relationships, most usually distances and/or angles, between such features in a molecule as atoms, ring centroids, and planes. The two 3D screens examined in this work were taken from the Tripos Unity 3D substructure searching system.

To produce a bit-string, a distance- or angle-range is first defined (e.g., 2−10 Å or 0−180°). This is then divided into a number of bins by specifying the bin width (e.g., 0.5 Å) producing 16 bins in the case of the example distances given here, when the first bin records all distances less than the minimum and the last all distances greater than the maximum. Pairs of features between which the distance or angle is to be indexed are defined, and each allocated a discrete set of bits in the bit-string corresponding to the number of bins. So, for example, the first 16 bits in a bit-string might index the distance between all nitrogen atoms and all oxygen atoms, individual bits being set if an N−O distance corresponding to that bin is present in the molecule. Bits 17−32 might index the distance between each N atom and each phenyl ring centroid and so on.

The Unity system produces two types of 3D screen. *Rigid* screens are based on distances measured in the single conformation. For our purposes this was the conformation generated by CONCORD[27] from the 2D structure. The Unity system default fingerprint that we used for our experiments indexes the distances between the following features in the range 2−10 Å in 0.5 Å increments: oxygen, nitrogen, generic oxygen or nitrogen, phenyl ring centroid, point on the normal to the plane of a phenyl ring and carbonyl extension point. This produces a string of 336 bits.

Flexible screens record all possible distances that a pair of features can achieve, based on the incremental rotation of all the rotatable bonds between the two features. No energy calculation is considered so some possible distances recorded in the screens may represent highly strained structures, and some combinations of individually allowed distances may be geometrically impossible. The system default fingerprint, which produces a string of length 240, indexes the distances between pairs of nitrogen, oxygen, or generic nitrogen or oxygen. Since this proved to be a very poor descriptor and for the sake of a direct comparison, the default definition of the rigid fingerprints described in the previous paragraph was also used as input to the flexible fingerprint generator. Each of these fingerprints took considerably longer to compute than the standard flexible fingerprints since many more distances are encoded.

Potential-Pharmacophore-Point Distance and Triangle Descriptors. The Unity 3D descriptors are based on encoding the types of atoms found in fragments or at points in space. They do not distinguish the behavior of atoms and would, for example, consider the nitrogen of a guanidium and nitro to be equivalent. Two descriptors examined in this study, which have been developed in-house, instead encode structures in terms of potential macromolecular recognition sites, or *pharmacophore points*, in an attempt to distinguish how atoms might behave. This precedent for this type of encoding is in the 3D-Search program of Sheridan *et al.*[26]

These descriptors are generated by a program known as 3D-FEATURES. This identifies all atoms in an input structure that could potentially act as hydrogen bond donors or acceptors and sites of potential positive or negative charge. The program also recognizes hydrophobic atoms. It represents connected groups of these of user defined size as a single geometric centroid, based on the CONCORD coordinates generated from the input structure.

3D-FEATURES contains an expert system with approximately 250 rules encoded in the Daylight SMARTS language.[23] These rules contain definitions of various 2D environments. Nesting of the rules up to seven levels relates definitions of specific atoms in a given environment to a classification of one or more types of pharmacophore point. An example is given in Figure 1 in which the various forms of enolate oxygen are encoded as potential sites of negative charge interaction. First the various forms of enolate oxygen (ONEENOLAT, TWOENOLAT, and THREEENOLAT) are defined, and these are set to be equivalent (ENOLAT). The enolate definitions are then brought together with other types of negative oxygen (ONEG), such the oxygen of acid groups (ACIDO). Negative oxygens are defined to be equivalent to other atom types with potential negative charge (NEG). Finally a single point NEG is defined such that all atoms that are included in the NEG definition will be identified and marked in all structures input to the program.

3D-FEATURES addresses two specific problems inherent in this type of classification system; to ensure that all protonation states and major tautomers are taken into account when structures are encoded. For example, at pH 7 the protonated state of the structure in Figure 2 will exist alongside the neutral form that would be more usually presented to the program. 3D-FEATURES contains rules defining these two nitrogens to be equivalent, at least for the purposes of charge interaction, and the atom will be labeled accordingly irrespective of which form is presented to the program. Tautomers are handled in a similar fashion. 3D-FEATURES assigns behavior according to all major tautomers. In the example in Figure 3, forms B and C are the major tautomers of the structure, and so the oxygen atom will be labeled as an acceptor and the two nitrogens as acceptors or donors, even if the structure is input in form A. Once again, one set of rules will define all three environments of the oxygen to be equivalent, and other sets will define the nitrogens to be equivalent.

There are two major difficulties with this type of system. One is accurately encoding behavior since there may be more than one opinion about the behavior of particular groups. The knowledge in our system has come from the experience of a number of Abbott and consultant chemists as well as a number of standard references.[28−30] The second problem, as with most expert systems, is to assess the completeness of coverage. In an ongoing process, the assignments of several thousand structures have been manually examined in an attempt to identify and add groups not covered by the rule base.

Having reduced a structure to its pharmacophore points, 3D-FEATURES then makes use of the CONCORD coordinates to calculate the distances between all such points.

These distances are used as the basis for two descriptors *potential pharmacophore point pairs* (PPP Pairs) and *potential pharmacophore point triangles* (PPP triangles).

PPP-Pairs. As its name suggests, this descriptor directly encodes the distance information between all pairs of potential pharmacophore points. A number of alternatives have been implemented for encoding the PPP-pair information into a bit-string suitable for similarity calculations.

The simplest uses the Unity 3D encoding style, allocating each of the 15 combinations of pairs (donor−donor, donor−acceptor, *etc.*) of the five point types a separate area of a bit-string and dividing each section into a number of bits based on a user specified minimum, maximum, and bin-width.

One variation on this method is to allow an overlap of the bins to be specified. This addresses one shortcoming of the standard method, that two very close distances that happen to fall either side of a bin boundary will not contribute to the similarity value. In our overlap encoding system two bits are used for each bin. The first is set according to the exact distance measured. If the distance does not fall within a user specified distance of the bin boundary the second bit is set. If it does, a bit is set in the bin on the opposite side of the boundary. To avoid biasing the similarity calculation all distances must set the same number of bits, hence the use of the second bit. The overlap is specified by the user as a percentage of the bin width.

A second variation substitutes bins of "equifrequent" occurrence for fixed width bins. This method is based on the substructure search keys used in the 3DSEARCH system of Sheridan *et al.*;[26] the same method is used in the Chem-X software.[25] Sheridan *et al.* examined the frequency distribution of interatomic distances in various 3D databases and noticed a peak at around 3−5 Å. Consequently they used narrow bins in this region and wider bins at higher distances to even out the distribution of the number of distance falling into each bin. We found that defining the bit position of a bin as follows gave a suitable distribution.

$$Bin\_Number = (\text{int})\left(5 \tan^{-1}\left(\frac{D-3}{2}\right) + 6\right)$$

D = distance between features

$\tan^{-1}$ is measured in radians

PPP-Triangles. The PPP-triangle descriptor encodes all potential three-point pharmacophores present in each molecule. Similarity calculations based on triangular descriptors have previously been described by other groups[31−33] although our encoding scheme is somewhat different from any of these. This scheme follows the same principle as before, making use of a minimum, maximum, and bin width to encode each distance.

There are 35 possible combinations of any three of the five pharmacophore point types (donor−donor−donor, donor−donor−acceptor, etc.). Each of these combinations is allocated a section in a bit-string. Each section is divided into sufficient bins to represent all combinations of the distances along each of the three edges given the minimum, maximum, and bin-width that are allowed by the triangle inequality. Using typical parameters of 2 Å, 15 Å, and 1 Å the first bin would represent (<2, <2, <2) and subsequent bins (<2, <2, 2−2.99), (<2, <2, 3−3.99), ..., (14−14.99, >=15, >=15), (>=15, >=15, >=15). A graphical representation of a single bin, HBD-HBA-HYD (6.00−6.99, 3.00−3.99, 5.00−5.99), is shown in Figure 4.

Using the parameters given above there are over 48 000 possible three point pharmacophores. It is therefore impractical to represent each bin with a bit in a bit-string. Instead, a hashing scheme similar to that used by Daylight and Tripos for their 2D fingerprints is used, in which each bin is randomly assigned a number of bits in a much smaller bit string. Setting five bits in a 2048 bit string was found by experiment to be suitable. The likelihood for collisions is very low, although nonzero, since the number of three-point pharmacophores in each molecule is very small.

CLUSTERING METHODS

The clustering methods that we examined may be divided into two main types, *nonhierarchical* and *hierarchical.* They are all nonoverlapping and therefore no structure appears in more than one cluster.

Nonhierarchical methods produce a single set of clusters with no relationships between those clusters. We used the method due to Jarvis and Patrick[34] as implemented in the Daylight clustering suite of programs.[24] Previous studies at Sheffield University[35] suggested this method produces a clustering in which structures of similar biological activity tend to group together. The method has become widely used for various diversity related tasks.

Jarvis−Patrick clustering functions in two stages. In the first, the similarity of every structure in the dataset to every other is calculated. Following the Sheffield study,[35,36] typically the Tanimoto coefficient of similarity is used which is a measure of the number of set bits in each string in common. For two structures i and j each characterized by a k-bit string, the Tanimoto coefficient is calculated as follows:

$$\text{Tanimoto} = \frac{\sum_k (x_{ik} x_{jk})}{\sum_k ({x_{ik}}^2 + {x_{jk}}^2 - (x_{ik} x_{jk}))}$$

In this *nearest-neighbor* stage the top n most similar structures to each structure in turn are recorded. n is a user-defined parameter; in our experiments the Daylight default value of 16 was used. In an addition to the Daylight programs, we implemented the use of a threshold on the nearest-neighbor calculation. This allows a structure to be a member of another's nearest-neighbor list only if the two have a greater than threshold similarity. This addition prevents a structure that has little similarity to any other members of the dataset producing a list containing a set of very dissimilar structures.

The second stage of the clustering uses the nearest-neighbor lists to assign compounds to clusters. Two structures A and B cluster together if

- A is in the top K nearest-neighbor list of B
- B is in the top K nearest-neighbor list of A
- A and B have at least K_{min} of their top K nearest neighbors in common, where K_{min} is a user-defined parameter less than K.

The number of clusters produced is governed by the choice of K and K_{min} and cannot be directly controlled. The major

advantage of Jarvis−Patrick clustering is that it is very fast; it has a time requirement of $O(N^2)$. One of its major disadvantages is that it has a tendency to produce a very high proportion of singletons under "strict" clustering conditions or a few very large clusters containing very diverse structures when using less strict conditions. Whilst the imposition of a threshold will help to prevent the latter problem, it will contribute to the former since it results in a decrease in the number of structures in nearest neighbor lists, particularly in sparse areas of space. This may result in two structures, which are very similar to each other, not having enough other neighbors in common to be able to cluster together and therefore becoming singletons.

A version of Jarvis−Patrick from the BCI clustering package[37] with an enhancement aimed at overcoming some of these problems was included in our study. As well as allowing a similarity threshold to be imposed, this implements variable-length rather than fixed-length nearest-neighbor lists. This allows *all* structures with a similarity greater than or equal to the threshold to be included in a nearest-neighbor list, irrespective of the length of the resulting list. At the clustering stage, the third condition for clustering is replaced with the requirement that A and B have at least P_{min} of the neighbor list in common, where P_{min} is a percentage of the length of the shorter list. In this case the two similar structures in a sparse area of space will be permitted to cluster, since they will have a high proportion of their short nearest neighbor lists in common; in the extreme case this will be 100% of lists consisting solely of the other structure. Variable-length nearest neighbor lists address another deficiency of the standard scheme. Nearest neighbors of equal similarity to others already included in a list may be excluded if the list is already "full". This may lead to the arbitrary splitting of large clusters of similar objects, particularly in dense areas of space. With variable length lists, all such neighbors will be included, and this problem will not arise.[38]

Hierarchical clustering methods produce classifications in which very small clusters of very similar structures are nested within larger clusters containing more diverse structures. A cluster hierarchy is often visualized as a dendogram in which at one extreme all clusters are singletons and at the other all structures are in a single cluster. Hierarchical clustering methods are said to be *agglomerative* if they start from singletons and work toward a single cluster or *divisive* if they work in the reverse directions. Hierarchical clustering algorithms are discussed in detail by Murtagh.[39]

Agglomerative methods may be described by the following process.

(1) Calculate the similarity of each structure in the dataset with all others.

(2) Find, and merge into a single cluster, the two most similar remaining objects (clusters or structures) in the dataset.

(3) If more than one object remains go to step 2.

We made use of two agglomerative methods in the BCI clustering package,[37] that differ only in the way in which they decide which two remaining objects are most similar. If the algorithm uses Euclidean distances in its calculations, then the program produces a *Ward's* hierarchy.[40] If instead the cosine similarity coefficient is used then a *group-average* hierarchy results. The two coefficients are calculated as follows:

$$\text{Euclidean} = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

$$\text{Cosine} = \frac{\sum_k (x_{ik} x_{jk})}{\sqrt{\left(\sum_k x_{ik}^2 \sum_k x_{jk}^2\right)}}$$

When merging two clusters, Ward's method has the effect of maximizing the intercluster variance whilst minimizing the intracluster variance. Group-average merges two clusters such that each cluster has a lesser average distance to the remaining members of that cluster than it does to all members of any other cluster.

If implemented as described above, the algorithm would have a time-requirement of $O(N^3)$. However, the BCI implementation makes use of a much faster method based on the *reciprocal nearest-neighbor* (RNN) algorithm of Murtagh.[39] The method is discussed in detail by Downs and Willett[15] and is of order $O(N^2)$.

We also used one hierarchical divisive method, the *Guénoche*,[41] or minimum-diameter method, as implemented in the BCI software.[37] This functions by recursively dividing the largest diameter cluster remaining in the hierarchy into two such that the larger resulting cluster has the smallest possible diameter. The algorithm functions as follows:

(1) Calculate the $n(n-1)/2$ dissimilarities between all n structures in a dataset, recording this in a list sorted in decreasing dissimilarity.

(2) Taking the pair of most dissimilar structures as the focus for the first partition, assign all other structure to the least dissimilar of these initial cluster centers.

(3) Recursively select the cluster with the largest diameter cluster and partition it into two such that the largest resulting cluster has the smallest possible diameter.

(4) Repeat step 3 for a maximum of $n - 1$ iterations.

At worst Guénoche will have a time requirement of $O(N^2 \log N)$ and a storage requirement of $O(N^2)$. For step 1 any measure that is embeddable in Euclidean space may be used. We considered both dissimilarity, calculated as (*1* − Tanimoto), and Euclidean distance.

EXPERIMENTAL DESIGN

Our purpose is to identify combinations of descriptors and clustering methods that allow the selection of representatives from each cluster. To be suitable such methods should cluster together biologically similar structures and more importantly separate actives and inactives into different sets of clusters. The first condition will ensure that the selection samples the range of activities in the dataset. The second minimizes the chance that any type of activity is missed when an inactive is selected as representative of a cluster containing any actives.

Our evaluation therefore consisted of producing clusters from each combination of a descriptor and a clustering method using a number of sets of structures of known activity and examining the numbers of actives and inactives within every cluster, to determine the degree of separation of one from the other.

The datasets used in this study were as follows:

•Monoamine Oxidase (MAO)—This set contains 1650 structures with 290 actives. These structures were predominately manually selected from existing Abbott compounds to find a new lead. Thus, this set has a relatively high diversity.

•Proprietary Enzyme Assay 1 (PEA1)—This set contains 2730 structures with 648 actives. PEA1 contains more synthetic series than the MAO dataset and so has a lower diversity than that dataset.

•Proprietary Enzyme Assay 2 (PEA2)—This set contains 800 structures with 379 actives. The structures in PEA2 are mostly from a small number of synthetic series and so have the lowest diversity and also have a much higher hit rate than the other sets.

•High Throughput Screening (HiTS)—A set of over 16 000 structures was available together with activities in a number of high throughput screens.

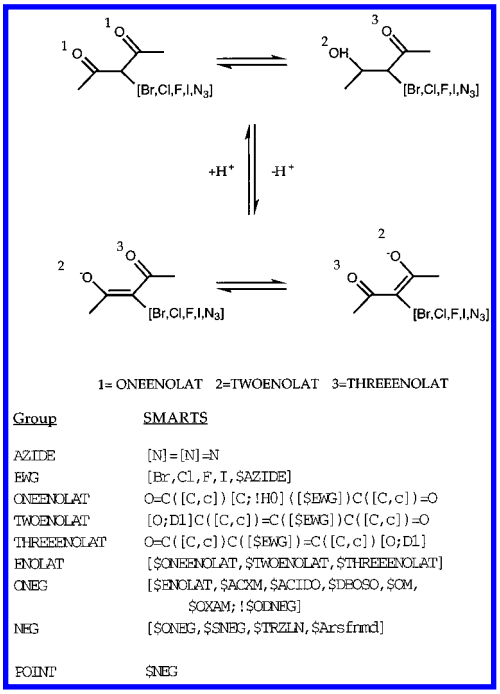
In the cases of MAO and HiTS the assay results directly classified the structures as active or inactive. For the two PEA sets, the active/inactive classification was assigned on the basis of IC_{50} values.

For every combination of one dataset, one clustering method and one descriptor, a number of sets of clusters were produced, either by sampling the hierarchies at regular intervals or by varying the parameters to the Jarvis–Patrick method. For the former the hierarchy was sampled at every 100 clusters for MAO and PEA2, every 250 clusters for PEA1, and every 1000 clusters for the HiTS data. For the latter, thresholds of 0, 0.6, 0.7, and 0.8 were used in combination with K_{min} values of 6, 8, and 10 for the standard version and thresholds of 0.6, 0.7, 0.8, and 0.9 in combination with P_{min} values of 50, 60, 70, 80, and 90% for the enhanced version.

The purpose of the analysis carried out on each of these sets of clusters was to determine the degree of separation of actives from inactives. If an *active cluster* is defined as one in which at least one member of a cluster is active, then a subset of the dataset, termed the *active cluster subset*, may be defined as the set of structures in active clusters. If the proportion of structures in this subset that are active (p_a) is compared to the proportion of actives in the dataset as a whole (p_0), the difference, if any, is an indication of the degree of separation of actives from inactives in that set of clusters.

Active singletons have to be excluded from this calculation, since individually they give a proportion of 1.0 and skew the results, especially in cases were there are a very large number of singletons. Singletons are undesirable since they are not representative of anything else, and, when considering compound selection, must all be selected and are not predictive of any other actives if found to be a hit.

An increase in the proportion of actives in the active cluster subset can arise in two ways. The first is simply as a result of the clustering. Whenever a dataset is partitioned into any number of clusters greater than the number of active structures, the actives may distribute at no more than one per cluster and inactive clusters will still be formed, since there will not be sufficient actives to "go around". Whenever this occurs, the proportion of actives in the active clusters will increase over the population as a whole, although no grouping of actives has occurred. This effect eliminates the bulk of the inactives which are not similar to any actives.



Group	SMARTS
AZIDE	[N]=[N]=N
EWG	[Br,Cl,F,I,$AZIDE]
ONEENOLAT	O=C([C,c])[C;!H0]([$EWG])C([C,c])=O
TWOENOLAT	[O;D1]C([C,c])=C([$EWG])C([C,c])=O
THREEENOLAT	O=C([C,c])C([$EWG])=C([C,c])[O;D1]
ENOLAT	[$ONEENOLAT,$TWOENOLAT,$THREEENOLAT]
ONEG	[$ENOLAT,$ACXM,$ACIDO,$PHOSO,$SOM, $OXAM;!$ODNEG]
NEG	[$ONEG,$SNEG,$TRZLN,$Arsfnmd]
POINT	$NEG

Figure 1. Daylight SMARTS definitions for the oxygen of the various forms of enolate shown in the structure diagrams, defining them to be potential negative charge pharmacophore points. (Note EWG = electron withdrawing group.)

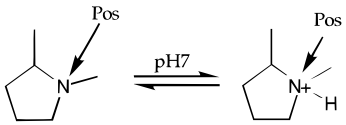


Figure 2. Assigning pharmacophore points according to all ionization states at pH7.

Clustering methods which do not form clusters containing many diverse structures will perform best in this respect.

A further increase in proportion will occur if there is any greater similarity between pairs of actives than active–inactive pairs. This depends both on the presence of such pairs in a dataset and the ability of a given descriptor to characterize this similarity. The extent to which this is occurring can be identified by repeating the analysis described above on the same sets of structures but with each randomly assigned to be active or inactive in a proportion mimicking the true activity data. In our experiments 20 sets of pseudoactivity data were generated in this way for each of the real sets, and all clusters analyzed against every one of these. Any increase in the proportion of actives in the active cluster subset when using the "real" data over all the random data indicates that the separation is due the presence of active–active pairs which are more similar to active–inactive pairs and that the descriptor is able to characterize the former as more similar to each other than the latter.

RESULTS

The discussions in this section concern the smaller MAO, PEA1, and PEA2 datasets.

Real vs Random Activity Data. Figure 5a shows the analysis of real and randomized MAO data against clusters generated using MACCS descriptors and Ward's clustering. The hierarchy is sampled at every 100 clusters, each sample producing a single point for the real activity data and 20 points for the randomized data. The *x*-axis plots the number

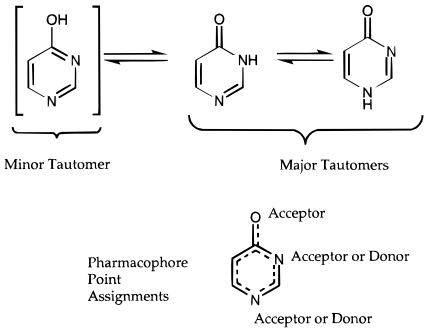


Figure 3. Assigning potential pharmacophore points according to major tautomers

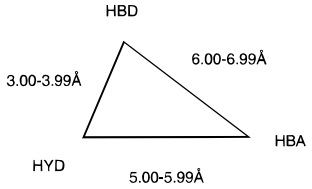


Figure 4. Definition of a PPP triangle descriptor bin at a tolerance of 1 Å.

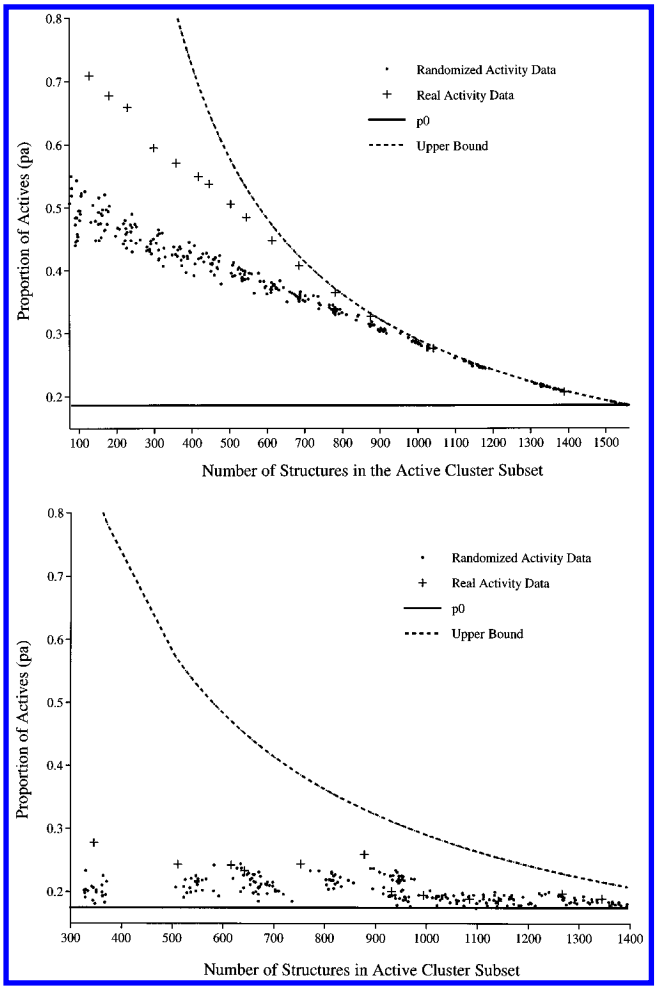


Figure 5. Analysis of real vs randomized MAO activity data using (a) MACCS descriptors and Ward's clustering and (b) Unity 3D flexible descriptors and Jarvis−Patrick clustering.

of structures in the active cluster subset, whilst the *y*-axis plots p_a, the proportion of those structures which are active. Moving from right to left across the graph corresponds to a tightening of the clustering conditions and an increase in the total number of clusters. Since the MAO dataset has 290 actives, primary consideration should be given to the portion of the graph between active cluster subset sizes of 400−600 when the greatest separation of actives from inactives will be achieved whilst most of the actives are in that subset. For example, at 400 clusters 220 of the 290 actives are in the multimember clusters rather than singletons. An upper bound to p_a is shown on the graph. This is calculated as

$$p_{a(\text{upper})} = \frac{\text{no. of actives in dataset (i.e., 290)}}{\text{no. of structures in active cluster subset}}$$

This is the maximum possible proportion of actives for a given size of active cluster subset. This would be obtained only if all active structures were in multimember clusters and none in singletons. When the size of the active cluster subset equals the total number of actives, $p_{a(\text{upper})} = 1.0$. This represents the ideal result in which all actives would be in multimember clusters, none of which would contain any inactives.

Comparison of the random points against the p_0 baseline shows that there is an enrichment of actives in the active clusters solely as a result of the partitioning, the effect increasing as the number of clusters increases. However, there is a much greater increase in proportion when analyzing the real data which shows that there is a further separation as a result of the actives being grouped together and therefore having a higher similarity to other actives than to inactives.

Figure 5b shows the equivalent result for the MAO dataset using standard Jarvis−Patrick clustering and Unity 3D flexible fingerprints generated using the Tripos default flexible fingerprint definition. In this case the majority of the data points from the real activity data are not separated from those calculated from the random data, indicating that this combination of methods has not been able to distinguish active from inactive. The only points that do produce a separation arise from the use of $k_{\min}/k = 12/16$. This produces extremely tight clusters and a very high proportion of singletons. Furthermore, many values of p_a are little better than the p_0 baseline, indicating that the clustering itself was unable to achieve any separation.

The only cases in which the later result was observed was with the Jarvis−Patrick clustering using Unity 3D flexible fingerprints. Partly for this reason, and partly to allow a direct comparison with the rigid fingerprints, all analyses described in the following section are based on using the Tripos 3D rigid default definitions to calculate flexible fingerprints.

For all combinations of descriptor and clustering method except the one mentioned above there was a clear separation of real from random data, and of random data from the baseline, of the type shown in Figure 5a.

Optimizing Descriptors. Direct comparisons between the various descriptors have been made by comparing clusters from the same dataset using the same clustering method.

MACCS Occurrence Counts. When using a structural keys based descriptor, it is possible to record the frequency of occurrence of each fragment instead of a simple presence or absence. The effect of using a varying occurrence count with the MACCS descriptors is shown in Figure 6; the descriptors encoding 1, 3, or 13 occurrences. The results, for PEA1 using Ward's clustering, show that there is a slight improvement using three occurrences rather than one but that

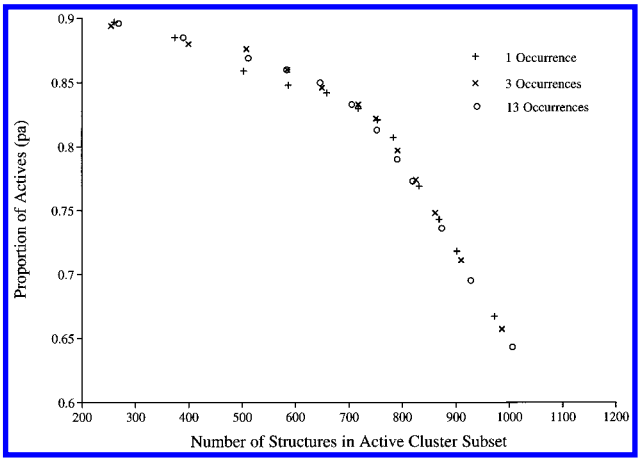


Figure 6. Analysis of PEA1 activity data using Ward's clustering varying the number of recorded occurrences for each fragment in the MACCS set.

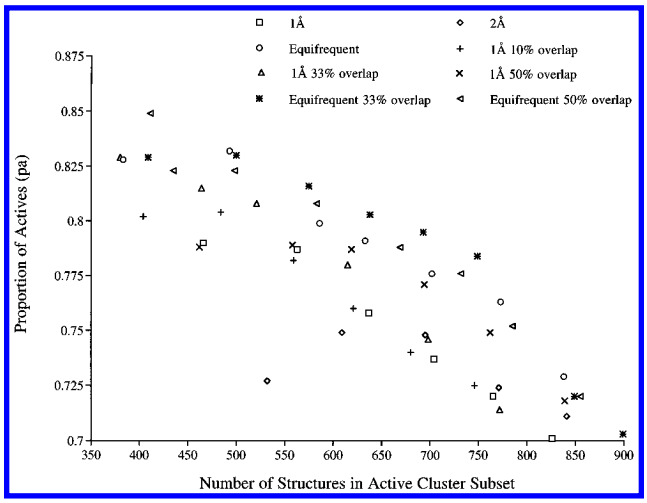


Figure 7. Analysis of PEA1 activity data using Ward's clustering and a variety of encoding schemes for the PPP-Pair fingerprints.

there is no clear advantage to using any more than three. Three was therefore used as the default for the remainder of this work.

Potential Pharmacophore Point Pair Fingerprints. A number of options were implemented for encoding the PPP-pair fingerprints. For each dataset a comparison was made using 1 and 2 Å bins and equifrequently populated bins with no overlap, 1 Å bins with 10%, 33%, and 50% overlap and equifrequent bins with 33% and 50% overlap. The results for dataset PEA1 using Ward's clustering are shown in Figure 7. These results, which are typical of those across all datasets, show that there is some difference between the utility of the different encoding schemes.

The most successful encoding schemes use approximately equifrequently populated bins. Of the fixed width bining schemes, a 1 Å bin width is better than 2 Å. For both equifrequent and fixed width 1 Å bins there is an advantage to using an overlap providing it is large enough. Overlaps 33% and 50% give similar results, 10% overlap seems to offer little advantage. The advantage to using an overlap appears to be greater for the fixed-width bins than equifrequent. For the remainder of this work all PPP pair descriptors were calculated using equifrequent bins and a 33% overlap.

Potential Pharmacophore Point Triangle Fingerprints. A more limited set of encoding schemes were implemented for the PPP-triangle descriptors. Of the possible bin-widths 1 Å tolerance on each edge of the triangle gave the best results. Values of 2 and 15 Å were used for the minimum and maximum, respectively.

Table 1. Mean CPU Time per Structure for Calculation of Descriptors from the PEA1 Dataset[a]

descriptor	CPU s/structure
Daylight	0.04
Unity 2D	0.19
MACCS	0.41
Unity 3D Rigid[a]	0.02
Unity 3D Flex[a]	10.10
PPP Pairs[b] (Equifreq 33% overlap)	0.37
PPP Triangles[b] (1 Å bins)	0.74

[a] Excludes time for 3D structure calculation (typically 0.05 s/structure). [b] Includes time for 3D structure calculation. Note that no exact time measurement was available for the SSKEYS but that their calculation required on the order of 1 CPUs per structure.

COMPARISON OF DESCRIPTORS

Calculation Times. The CPU times per structure required to calculate each descriptor are shown in Table 1. These are measured for the PEA1 dataset using an Silicon Graphics Challenge (R8000). All but the flexible fingerprints require less than three-quarters of a second per structure to calculate, and so it is quite feasible to calculate these descriptors for very large datasets in a reasonable amount of time. The Unity flexible fingerprints, using the Tripos standard rigid definition, required over 10 s per structure to calculate making them less useful for very large databases.

Grouping of Actives. Figure 8a−c shows the proportion of active structures in the active cluster subset, p_a, plotted against the size of that subset for datasets MAO, PEA1, and PEA2, respectively. The results were generated using Ward's clustering; equivalent trends were observed for other clustering methods.

All the 2D descriptors show a higher performance than any of the 3D ones across all the datasets. Furthermore, the outcomes using any of the 2D descriptors is very similar. For MAO and PEA2 the MACCS fingerprints are marginally superior to the SSKEYS or hashed fingerprints; for PEA1 they are equivalent to the Unity 2D and SSKEY fingerprints. There appears to be little advantage to the Unity scheme of splitting the information from separate length paths into different areas of the bit-string, compared to the Daylight scheme in which all path information is hashed over the whole bit-string. Somewhat more surprisingly, there also appears to be no advantage to encoding the much greater amount of information in the hashed fingerprints as opposed to the structural keys, at least when this larger amount of information has to be overlaid in a bit-string.

Of the 3D descriptors the PPP-pairs produce the best results. For PEA2 they are as effective as some of the 2D descriptors, for MAO and PEA1 they do not reach this level. There appears to be no advantage to encoding triangles of potential pharmacophore points as opposed to pairs, the former always produce poorer results than the latter. This may be a result of the hash encoding scheme that had to be used to produce bit-strings from the triangles, in which information from different triangles could possibly become overlaid.

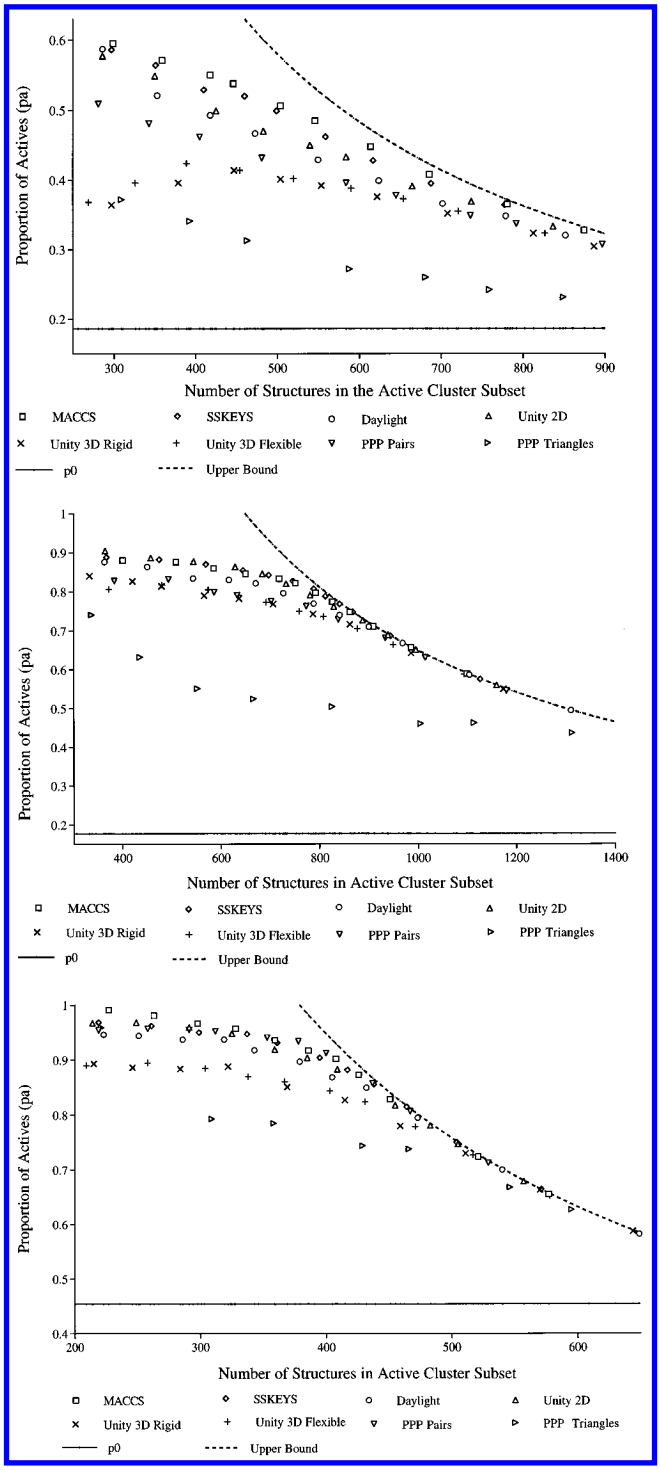


Figure 8. Analysis of (a) MAO, (b) PEA1, and (c) PEA2 activity data using Ward's clustering and all descriptors.

The two Unity 3D fingerprints are poorer than the PPP-pairs for all datasets except for PEA1 in which the performance is approximately equal. There seems to be very little advantage in using the flexible fingerprints rather than rigid ones, if the same features are indexed in both. There is however a very large time penalty when calculating the flexible fingerprints; on average they required 500 times as much CPU per structure to calculate as the rigid ones. As discussed previously, when using the standard flexible fingerprint definition, which contains many less features and so is faster to calculate, the results were considerably worse than for any other descriptors across all the datasets with any of the clustering methods.

Table 2. CPU s Required To Cluster Dataset MAO of 1650 Structures Using MACCS Descriptors

method/dataset	step 1	step 2	total
Ward	66.3	1.4	67.7
group average	92.0	1.4	93.4
Guénoche	28.3	10.1	38.4
Jarvis−Patrick	4.1	2.3	6.4

[a] Step 1: •Ward, group average: RNN calculation; •Guénoche: produce sorted list of distances; •Jarvis−Patrick: nearest neighbor calculation, add 0.7 threshold. Step 2: •Ward, group average: extract 750 clusters; •Guénoche: construct hierarchy and extract 750 clusters; •Jarvis−Patrick: extract clusters with $k_{min} = 8/16$.

COMPARISON OF CLUSTERING METHODS

The comparisons described in this section are based on the use of MACCS keys since these were shown to be most successful in the experiments described in the previous section. However, the trends observed using the MACCS descriptors were very similar with all the other descriptors.

CPU Requirements. The CPU times required for each method to produce approximately equivalent sets of clusters are shown in Table 2. The agglomerative methods are shown to require the largest CPU time. Jarvis−Patrick clustering is by far the fastest of the methods considered. Timings were not available for the enhanced version of Jarvis−Patrick; however, it differs primarily in the length of the lists written and not in the computation required and so should have a similar CPU requirement to the standard method.

Grouping of Actives. The graphs in Figure 9a−c show the results obtained for datasets MAO, PEA1, and PEA2, respectively. In all cases the most successful method is Ward's. Group-average and Guénoche are then approximately equal in performance. Except in the case of PEA2 there seems to be little difference between the results from the Guénoche clusters computed using Euclidean or Tanimoto distances. In all cases standard Jarvis−Patrick is the least successful method, and in the cases of PEA2 and MAO the results are considerably worse than those produced from any of the hierarchical methods. The poorest results from Jarvis−Patrick arise when no threshold is used. The results from these two datasets also show that there is less control possible with standard Jarvis−Patrick in terms of the number of clusters produced.

The differences between the methods is partly related to the evenness of the distribution of cluster sizes that they produce. With very tight clustering conditions, i.e., a high threshold and k_{min}, Jarvis−Patrick clustering produces a very high number of singletons. With less rigid conditions it tends to produce one or two very large clusters containing very diverse structures. In the former case, most actives will be singletons; in the latter, most will be in the large clusters along with many inactives.

In contrast, all the hierarchical methods produced a much more even distribution of cluster sizes. At levels of clustering in which all actives are still included in the active cluster subsets, these methods produce many small clusters (typical cluster sizes range between 2 and 8) and produce very good separations of active from inactive.

The enhanced version of Jarvis−Patrick is more successful than the standard version for the more diverse data in MAO. For the less diverse datasets there is less difference. For the more diverse data, the variable length lists will allow

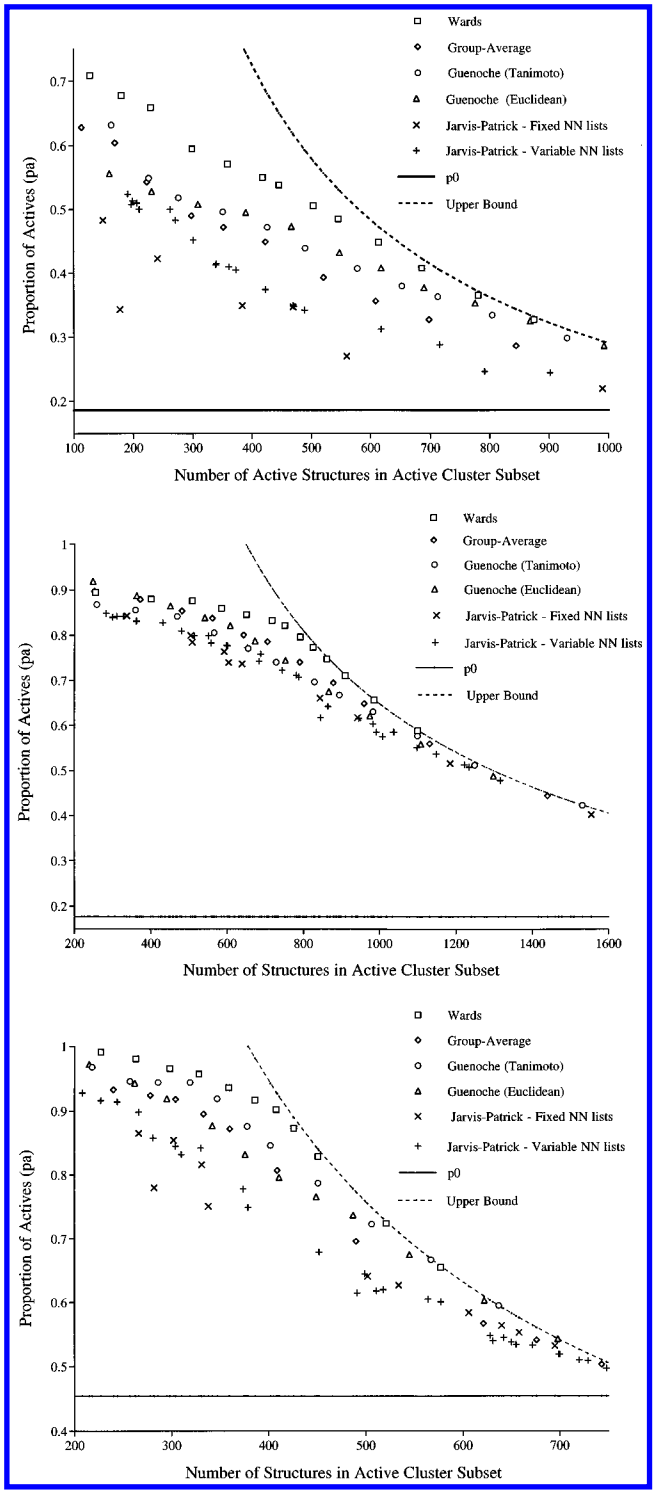


Figure 9. Analysis of (a) MAO, (b) PEA1, and (c) PEA2 activity data using MACCS descriptors and all clustering methods.

similar structures in the sparse areas of space in the MAO dataset to cluster together. For the more homogeneous datasets, the two sets of neighbor lists should be much more similar, since many more structures will have sufficient neighbors to make a "complete" fixed list and so the two methods perform similarly. Another advantage of the enhanced Jarvis−Patrick is apparent from the graphs, that it is easier to produce a continuum of sets of clusters, by continuously varying the threshold and P_{min} values. This provides more control when a particular number of clusters are required; with the standard implementation it is sometimes difficult to find conditions which will give close to a required number of clusters.

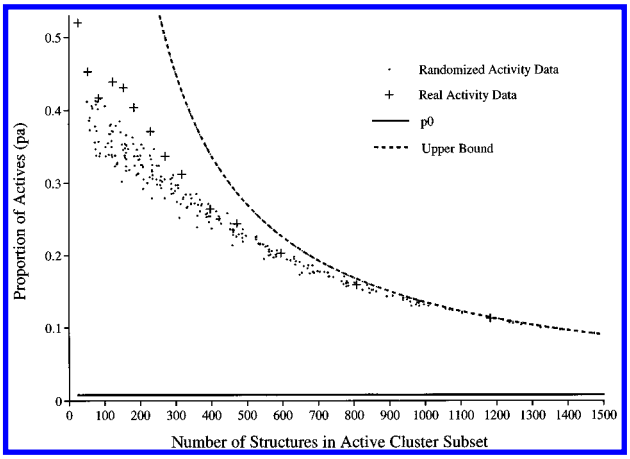


Figure 10. Analysis of real vs randomized HiTS data using Ward's clustering and Unity 2D fingerprints.

Table 3. Number of *n*th Most Similar Structures to Each Active Structure That Are Themselves Active

n	HiTS	PEA2
1	11 (8%)	215 (57%)
2	10 (7%)	203 (54%)
3	8 (6%)	186 (49%)
4	6 (4%)	200 (53%)
5	3 (2%)	185 (49%)
6	4 (3%)	201 (53%)
7	3 (2%)	191 (50%)
8	4 (3%)	185 (49%)
9	5 (4%)	173 (46%)
10	3 (2%)	177 (46%)

[a] In parentheses the number is given as a percentage of the total number of actives.

ANALYSIS OF HITS DATA

Analyses equivalent to those described above were conducted on various assays from the high-throughput screening program. The hit rates were much lower in these screens than any of the smaller datasets. The results presented in this section are from an assay with 135 hits in 16 380 tests, a 0.82% hit-rate. Similar results were found from other high-throughput assays with a similar hit-rate. In cases where there were only a handful of hits, no clear results were obtained.

Real vs Random Activity Data. Figure 10 shows the analysis of both real and random activity data analyzed against clusters generated using Unity 2D fingerprints and Ward's clustering. A considerable separation of actives from inactives is being achieved in this experiment; approximately 30−50% of the structures in the active clusters subset are active compared to 0.82% in the dataset as a whole. The graph shows that the separation is largely a consequence of the ability of this clustering method to produce only small, low diameter clusters; the *diameter* of a cluster is the distance between the two most dissimilar structures within that cluster. Hence, the clustering is not grouping actives together so much as separating off dissimilar inactive compounds.

To illustrate the reason for this, Table 3 shows the results of similarity searches, based on MACCS keys, for each of the actives in the HiTS dataset. The ten most similar compounds to each active were examined, and the numbers

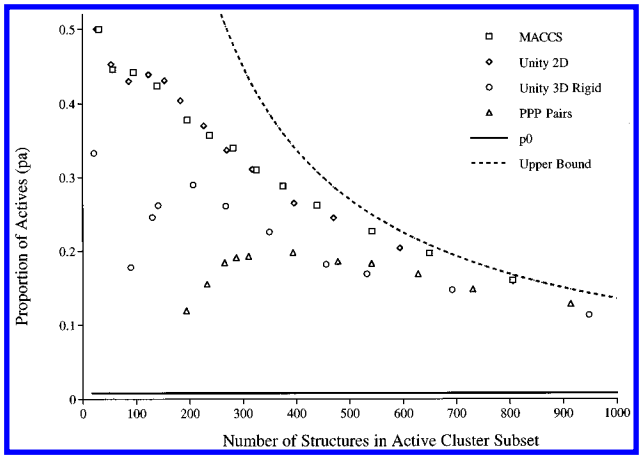


Figure 11. Analysis of HiTS data using Ward's clustering and a variety of structural descriptors.

Table 4. CPU Seconds Required for Clustering the HiTS Dataset of 16380 Structures[a]

method	step1	step2	total
Ward	5730.8	32.5	5763.3
group-average	10828.6	32.0	10860.6
Guénoche	1271.0	2057.3	3328.3
Jarvis−Patrick	704.5	19.2	723.7

[a] Steps are as in Table 2, except that 7500 clusters are extracted using the hierarchical methods.

of each of these which were also active are reported in the table. Equivalent results are given for the PEA2 dataset. The results show that the most similar structures to any HiTS active structure are predominately inactive, whereas the most similar structure to each PEA2 actives are much more likely to be active. For example, only 8% of the most similar structures to HiTS actives are themselves active; for PEA2 this figure is 57%. This is the reason that the clustering, in the HiTS case (unlike the PEA2 case), is unable to achieve much of an additional separation due to the clustering of actives together.

Comparison of Fingerprints and Clustering Methods. Figure 11 shows the performance of MACCS, Unity 2D and 3D rigid and PPP-pairs, using Ward's clustering against the HiTS data. As with the smaller datasets, there is a clear advantage to using the 2D rather than 3D fingerprints; however, there is little difference between the MACCS and Unity 2D fingerprints in this case. Since the descriptors' calculation methods are all of order O(N), calculation of the descriptors will require a similar mean CPU time per structure to that shown in Table 1.

Table 4 shows the CPU times required by the clustering methods. For an approximately tenfold increase in the number of structures from the times in Table 2, Ward's and Guénoche have an approximately 85-fold increase in CPU, group-average and standard Jarvis−Patrick approximately 115 fold.

Figure 12 shows the analysis of clusters produced using MACCS fingerprints and the various clustering methods against the HiTS data. There is a very clear difference between Ward's and Guénoche on the one hand, that produce a reasonable separation, and either implementation of Jarvis−Patrick on the other. Group-average produces a similar result to Ward's when the cluster size is small but similar to Jarvis−Patrick when it is not.

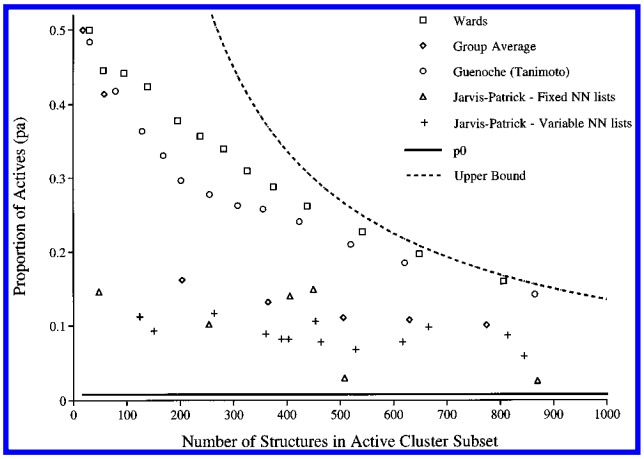


Figure 12. Analysis of HiTS data using MACCS descriptors and all clustering methods.

The outcome again seems related to the evenness of the distribution of cluster sizes and the ability of better methods to produce only small diameter clusters. Ward's and Guénoche produce a relatively even distribution irrespective of the level of the hierarchy examined. With more loose clustering conditions, group-average has the tendency to produce a few very large clusters containing many diverse structures, as does standard Jarvis−Patrick. As the clustering conditions tighten, group-average begins to produce a more even distribution; standard Jarvis−Patrick merely produces more singletons.

DISCUSSION

In this paper we have compared the use of a variety of structural descriptors and clustering algorithms with the intention of determining which are best able to separate known active structures from known inactives.

Our experiments suggest that the 2D structural descriptors available in the MACCS, Daylight, and Unity systems are better able to distinguish actives from inactives than any of the 3D descriptors we examined. Of the 2D descriptors, the structural keys are either equally effective or more effective than the hashed 2D fingerprints. This is somewhat surprising at first sight since the hashed schemes encode much more structural information. The results imply that the necessity to overlay this information in a relatively small number of bits might have a greater detrimental effect than previously imagined and, in the cases we examined at least, negate whatever advantage would be achieved by having the extra amount of information available. It might also be that some of the path information is redundant, with a detrimental effect on the similarity or distance calculation. The 153 key subset of the SSKEYS used in the MACCS descriptor appears as effective as using the whole set, given that an occurrence count of up to three was recorded for the latter.

The speed of the agglomerative clustering methods is related to the mean number of bits set in each descriptor. Hence, there is a further advantage to using the shorter and more sparse MACCS keys rather than the hashed fingerprints in terms of speed of clustering.

Although our results demonstrate that the sets of keys we investigated can distinguish active from inactive compounds, the results should not be construed to indicate that these keys would be the optimal ones for an analysis of what distinguishes actives from inactives in a given dataset. Rather,

for each biological endpoint one could search for a particular set of keys for this purpose. However, these keys have been shown to be effective across a number of different biological activities and hence are a set suitable for selecting compounds for testing against multiple endpoints or when no actives are known.

A comparison of the 3D descriptors suggests that encoding a structure in terms of its potential noncovalent interactions is more effective in characterizing bioactivity than a description in terms of more arbitrary substructures. Since the PPP based fingerprints inherently contain much more information about the 2D environments of the points being indexed, this may be an important factor in this result. The observation that the default Unity flexible fingerprints, that encode virtually no 2D environment information, performed far less well than the same type of fingerprints calculated using the default rigid definitions seems to confirm this. The choice of encoding scheme for the PPP pairs was shown to make a difference to the effectiveness of the descriptor, with the use of equifrequently populated bins and an overlap proving most successful.

The results suggest that the 2D descriptors originally developed for substructure searching can be effective for use in similarity calculations to distinguish biological activity. The 3D substructure search screens, however, seem less appropriate for this application. When a particular problem requires the 3D structure of molecules to be taken into account, our results suggest that a more detailed descriptor such as the PPP pairs should provide better results. We anticipate an improved performance with a new version of the PPP descriptors, currently under development, that will include not only the locations of atoms but also projected receptor site-points.[42] A 3D descriptor explicitly designed for similarity calculations which would encompass the whole molecule might, however, prove more effective than any of the pharmacophore or fragment and distance based descriptors, if one could be produced which would allow pairwise similarities to be calculated sufficiently fast to cluster large datasets.

Comparisons among the clustering methods show Ward's hierarchical agglomerative method to consistently be best able to separate active and inactive structures. The performance of group-average and Guénoche hierarchical methods were broadly similar and only slightly worse than Ward's. In all cases standard Jarvis−Patrick was the poorest method and in three of the four datasets examined considerably so. This result is broadly in agreement with the findings of Downs *et al.*[43] who first showed the advantage of hierarchical over nonhierarchical clustering methods for chemical datasets, in their case when producing clusters based on property-based rather than structural descriptors. Standard Jarvis−Patrick also produced the most uneven cluster distribution sizes and was prone to producing very many singletons as well as large diverse clusters. The enhanced version of Jarvis−Patrick addresses these problems using variable-length nearest neighbor lists and produced better results than the standard implementation for the more diverse data.

The standard Jarvis−Patrick method has been widely used for structure-based clustering for compound selection. Our experiments suggest that the method may not be as suited to this application as any of the hierarchical methods since there will be a greater chance of selecting an inactive compound as representative of one or more actives.

The main advantage to Jarvis−Patrick clustering is its speed, which makes it feasible to cluster hundreds of thousands of structures, even when using variable-length nearest neighbor lists. In comparison the agglomerative methods are much slower. However, with the RNN implementation and a fast workstation it is feasible to process large datasets. Using Ward's, processing a dataset of 100 000 structures characterized by MACCS fingerprints required 54 h CPU using the SGI Challenge; with the much more sparse PPP-pairs fingerprints, it was approximately three times faster. The Guénoche method is limited more by its storage requirements than CPU, requiring $O(N^2)$ storage. Currently this limits the method to 16 000 structures although future implementations should allow 64 000 structures to be processed.

Our results suggest that to achieve a good separation of actives from inactives, it is necessary to use levels of the hierarchy for which the mean cluster size is relatively small, perhaps around five. This means that it is probably unreasonable to hope to characterize a large diverse dataset with a small set of representatives but rather that more than 20% of the dataset may have to be tested.

CONCLUSIONS

The results described in this paper indicate that it is reasonable to base a compound selection or testing strategy on the use of clustering, provided a careful choice is made of which descriptors and algorithms to use. Selection of representatives from all clusters should allow the range of diversity among active compounds to be sampled and the number of types of activity missed to be minimized. Furthermore, once a number of active compounds have been identified for a given screen, testing of all compounds that cluster with those hits should result in a large number of further hits being identified, allowing the widest possible choice of lead compounds for optimization.

There are a number of practical considerations to take into account when using these clustering techniques for compound selection problems. These include determining a cutoff value to the similarity or distance between two compounds such that they still have a meaningful biological similarity; deciding which levels of a cluster hierarchy are appropriate to use for the selection; determining the minimum number of structures required to represent a dataset; and comparing the diversity of datasets based on the clustering results. We will address these issues in a forthcoming paper.

ACKNOWLEDGMENT

We would like to thank the following people and companies for their advice and assistance in this work. Joe Donahue, Brian Herr, and Steve Muskal of MDL Information Systems, Inc. for access to the SSKEYS; John Barnard and Geoff Downs of Barnard Chemical Information, Ltd. for access to the variable-length nearest neighbor list version of Jarvis−Patrick; Jeremy Yang of Daylight Chemical Information Systems, Inc.; Tad Hurst of Tripos Associates, Inc.; Robert Pearlman and Renzo Balducci of University of Texas at Austin.

REFERENCES AND NOTES

(1) Willett, P.; Winterman, V.; Bawden, D. Implementation of Non-Hierarchic Cluster Analysis Methods in Chemical Information Sys-

tems: Selection of Compounds for Biological Testing and Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(2) Lajiness, M. S.; Johnson, M. A.; Maggoria, G. M. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Liss, A. R., Ed.; 1989; pp 173−176.
(3) Johnson, M.; Lajiness, M.; Maggoria, G. In *QSAR: Quantitative Structure−Activity Relationships in Drug Design*; Liss, A. R., Ed.; 1989; pp 167−171.
(4) Bawden, D. In *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988.
(5) Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Method on an Initial Sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66−71.
(6) Whaley, R.; Hodes, L. Clustering a Large Number of Compounds. 2. Using the Connection Machine. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 345−347.
(7) 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**, *In press*.
(8) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67.
(9) Hansch, C.; Unger, S. H.; Forsythe, A. B. Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents. *J. Med. Chem.* **1973**, *16*, 1212−1222.
(10) Johnson, M.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(11) Gordon, A. D. *Classification*; Chapman and Hall: London, 1981.
(12) Everitt, B. S. *Cluster Analysis*; Edward Arnold: London, 1993.
(13) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; W. H. Freeman: San Franciso, 1973.
(14) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of 2-D Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(15) Downs, G. M.; Willett, P. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1994; Vol. 3.
(16) Feldman, A.; Hodes, L. An Efficient Design for Chemical Structure Searching. I. The Screens. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 147−152.
(17) Hodes, L. Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 88−93.
(18) Lynch, M. F. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Ed.; Ellis Horwood: Chichester, 1975.
(19) *Maccs II*; Molecular Design Ltd,: 14600 Catalina St. San Leandro, CA 94577. (510)-895-1313.
(20) Hazen, G.; Mikesell, J.; Shier, R. *MACCSII. Facilities Guide and Reference*; Molecular Design Ltd.: San Leandro, CA, 1987.
(21) *Unity Chemical Information Software ver 2.3*; Tripos Associates: 1699 S Hanley Road, Suite 303, St. Louis, MO 63144. 1-800-323-2960, support@tripos.com.
(22) *Unity System Manual*; Tripos Associates: St. Louis, Mo, 1994.
(23) James, C. A.; Weininger, D. In *Daylight Software Manual, version 4.41*; Daylight Chemical Information Systems Inc.: Irvine, CA, 1995.
(24) *Daylight Chemical Information Software, ver 4.41*; Daylight Chemical Information, Inc.: 18500 Von Karman, #450, Irvine, CA, 714-476-0451, info@daylight.com.
(25) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(26) Sheridan, R. P.; Nilikantan, R.; Rusinko, A.; Bauman, N.; Haraki, K.; Ventataraghavan, R. 3DSEARCH: A System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255−260.
(27) *CONCORD, A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*;The University of Texas at Austin and Tripos Associates: St. Louis, MO.
(28) Abraham, M. H.; Duce, P. P.; Prior, D. V.; Derek G. Garratt; Morris, J. J.; Taylor, P. J. Hydrogen Bonding. Part 9. Solute Proton Donor and Proton Acceptor Scales for Use in Drug Design. *J. Chem. Soc., Perkin Trans. II* **1989**, 1355−1375.
(29) Abraham, M. H. Hydrogen Bonding. Part 10. A Scale of Solute Hydrogen-bond Basicity using logK Values for Complexation in Tetrachloromethane. *J. Chem. Soc., Perkin Trans. II* **1990**, 521−529.
(30) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa Predictions for Organic Acids and Bases*; Chapman and Hall: London, 1981.
(31) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.
(32) Bemis, G. W.; Kuntz, I. D. A. Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput-Aid. Mol. Des.* **1992**, *6*, 607−628.
(33) Nilikantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.
(34) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans Comput* **1973**, *C-22*, 1025−1034.
(35) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(36) Willett, P.; Winterman, V. A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct. Activ. Relat.* **1986**, *5*, 18−25.
(37) *BCI Clustering Package, versions 2.5 & 3.0*; Barnard Chemical Information, Ltd.: 46 Uppergate Road, Sheffield, S6 6BX UK. (+44)-114-233-3170, barnard@bci1.demon.co.uk.
(38) Barnard, J. M.; Downs, G. M. Jarvis−Patrick *Clustering Documentation*; Barnard Chemical Information, Ltd.: Sheffield, UK, 1995.
(39) Murtagh, F. *Multidimensional Clustering Algorithms. COMPSTAT Lectures, 4*; Physica-Verlag: Vienna, 1985.
(40) Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **1963**, *58*, 236−244.
(41) Guénoche, A.; Hansen, P.; Jaumard, B. Efficient Algorithms for Divisive Hierarchical Clustering with Diameter Criterion. *J. Classif.* **1991**, *8*, 5−30.
(42) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A Fast New Approach to Pharmacophore Mapping and its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput-Aid. Mol. Des.* **1993**, *7*, 83−102.
(43) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

CI9501047