Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering

D. J. Wild and C. J. Blankley*

Parke-Davis Pharmaceutical Research Division, Warner-Lambert Company, 2800 Plymouth Road, Ann Arbor, Michigan 48105

Received July 26, 1999

Four different two-dimensional fingerprint types (MACCS, Unity, BCI, and Daylight) and nine methods of selecting optimal cluster levels from the output of a hierarchical clustering algorithm were evaluated for their ability to select clusters that represent chemical series present in some typical examples of chemical compound data sets. The methods were evaluated using a Ward's clustering algorithm on subsets of the publicly available National Cancer Institute HIV data set, as well as with compounds from our corporate data set. We make a number of observations and recommendations about the choice of fingerprint type and cluster level selection methods for use in this type of clustering

INTRODUCTION

Cluster analysis techniques are designed to find groups, or clusters, in data sets, where each member of a cluster is similar to the other members of the cluster based on a given set of descriptors. They have been widely applied in chemical structure handling applications, particularly for finding clusters of compounds with similar structures or physical properties. With the advent of high-volume screening (HVS), where large and diverse collections of compounds can be screened for biological activity, cluster analysis is becoming an increasingly popular tool for partitioning data sets into structurally related groups for further analysis. Barnard and Downs[1] and Willett[2] have reviewed a number of types of clustering algorithm that can be used for chemical structure handling applications. These generally fall into one of two categories: hierarchical or nonhierarchical methods. Hierarchical methods either begin with each item in a cluster by itself, and then progressively merge clusters until all of the items are in a single cluster (hierarchical agglomerative), or they begin with all of the items in a single cluster, and then progressively divide clusters until all of the items are in clusters by themselves (hierarchical divisive). In both cases, a hierarchical tree structure is produced, and taking slices across the hierarchy at different levels yields different distinct clusterings of the data set. Nonhierarchical methods generally produce a single clustering of a data set without any hierarchical relationships.

Two methods have become especially popular for use in chemical structure handling applications: one hierarchical (Ward's[3]) and one nonhierarchical (Jarvis−Patrick[4]). Recent work has shown Ward's method to be better than Jarvis−Patrick for separating known actives and inactives[5] and for property prediction[6] and to be effective for the selection of a diverse set of compounds.[7] We have thus used Ward's clustering for the experiments in this study. Ward's is a hierarchical agglomerative method that initially places each of the points (compounds in this case) in their own cluster. The method iteratively merges the two clusters with the shortest Euclidean distance between their centroids (normalized for the cluster size), until all of the compounds have been merged into a single cluster. A hierarchy of cluster levels is thus produced.

The Brown and Martin study[5] evaluated the suitability of different fingerprint types and clustering methods for clustering active compounds together, thus addressing the question of which fingerprints encode information that is relevant for biological activity. This theme was pursued further in a second study by Brown and Martin[8] which assessed how different fingerprint descriptors encoded information that correlated with properties known to be relevant for binding (electrostatics, sterics, hydrophobics, hydrogen-bonding ability, and dispersion). In this study, our first objective was to consider the more basic question of how well the different fingerprint types represent structural information such that compounds within a chemical series are clustered together. This was motivated by a desire to be able to group the active compounds coming from our high-throughput screening experiments into series that could be subject to a structure−activity relationship analysis.

Both hierarchical and nonhierarchical methods require the selection of an appropriate clustering. In the case of hierarchical methods, this involves selection of a level from the hierarchy that is output by the clustering algorithm, while nonhierarchical methods require input parameters to be set, such as a set of initial cluster centroids (K-Means) or the stringency required for membership of a cluster (Jarvis−Patrick). Since there is no obvious way of determining these parameters or for selecting an appropriate level for a particular set of chemical structures, this presents a problem. Our second objective was thus to find a method of selecting a level from a Ward's hierarchy that effectively represents the chemical series present in the data set.

Many methods for evaluating the "goodness" of a particular cluster level have been proposed and evaluated, mainly in the psychology literature. Milligan and Cooper[9]

* To whom all correspondence should be addressed. Telephone: (734) 622-7733. E-mail: john.blankley@wl.com.

10.1021/ci990086j CCC: $19.00
Published on Web 11/23/1999

have reviewed the effectiveness of 30 such measures in finding the correct number of clusters in 50-point data sets containing between two and five distinct clusters. We evaluated nine measures and compared their effectiveness in selecting cluster levels from a Ward's clustering which cluster compounds of the same chemical series together.

FINGERPRINT TYPES

We used four widely used descriptors for our clustering experiments: MACCS keys, Daylight fingerprints, Unity fingerprints, and BCI fingerprints. All of them generate a bit string based on two-dimensional structural information.

MACCS keys[10] use a predefined dictionary of either 166 or 960 structural features and register their presence or absence in a compound by a 1 or 0 in the corresponding position in a bit string. MACCS keys are therefore, strictly, structural keys, and there is a one-to-one correspondence between bits and features. MDL have only made public the definitions for the 166 set of features, and we used this set for the experiments reported here. MACCS keys may be generated using the SSKEYS program distributed by MDL.

Daylight fingerprints[11] do not require a dictionary, but encode information about all of the atoms plus all the 1 to 4, 5, 6 or 7 bond-length linear structural paths present in a compound. This information is hashed down onto 2048 bits and optionally folded down to 1024 bits. There is no direct relationship between individual bits and any one structural feature. For our experiments, we allowed the encoding of two to seven bond-length fragments and used nonfolded 2048 bit fingerprints. We generated Daylight fingerprints using functions in the Daylight Toolkit.

Unity fingerprints[12] are similar to Daylight fingerprints, except that they segregate different length paths into different regions of the fingerprint and permit a limited degree of customization of the fingerprint. Specifically, a fingerprint may contain paths of a specified range of sizes present in a compound and also may encode for specific fragments encoded in Sybyl Line Notation (SLN). We used the default 2D fragment descriptor which contains 928 bits corresponding to the paths present in a compound, plus 60 bits which are keys which encode for common and rare atoms, generic atom types, and simple generic rings. Several bits are used for each key to enable counts of the number of occurrences to be kept. The total fingerprint length is 988 bits.

BCI fingerprints[13] are unique in that they allow the generation of user-defined dictionaries, either made from scratch or derived from the fragments present in a set of compounds, and then allow these dictionaries to be refined and manipulated flexibly. The BCI software allows the use of six families of fragments in the fingerprint: augmented atoms (atoms with their connected bonds); atom sequences (linear paths, default two to six atoms); atom pairs (two atoms and the topological distance between them); ring composition (sequence of atoms and bonds around a ring); ring fusion (ring connectivities around each ring); and ring ortho (stereo configuration of nonplanar pairs of ortho-fused rings). There is complete flexibility about which fragment types are used in the fingerprint. We used two fingerprint formats for our experiments. The first (which we shall refer to as BCI/D) uses a BCI default dictionary of 4096 fragments (all family types except ring ortho) that is supplied with the software. The second (which we shall refer to as BCI/G) used dictionaries generated specifically for each of the data sets tested, containing all of the fragments for all families present in the compounds comprising the data set. BCI also provides a program called Pickfrag which allows the interactive refinement of generated dictionaries, but in this work we were only concerned with methods which could be fully automated, so we did not employ Pickfrag. The BCI/D fingerprints contained 4096 bits (one for each dictionary fragment, making it technically a structural key), and the length of the BCI/G fingerprints is dependent on the data set used (one bit for each fragment).

CLUSTER LEVEL SELECTION METHODS

We chose nine cluster level selection methods to test. All but the Kelley method are reviewed in papers by Milligan and Cooper.[9,14−16] We selected methods on the basis of their simplicity of implementation and lack of need for parametrization.

A number of the methods that use a measure of mean or summed within-cluster distances are ambiguous in relation to singletons. Using a value of zero for singleton mean or summed within-cluster distance or disregarding them entirely would mean that cluster levels containing a large number of singletons would be favored over levels without singletons. For example, a cluster level consisting entirely of singletons would have a C-index value of zero and would thus be chosen as the optimal level. This problem was addressed by Milligan and Cooper[9] by only considering cluster levels where $(n_c + n_s) < n/2$ (n_c is the number of clusters, n_s is the number of singletons, and n is the number of points). We have adopted this approach in our experiments. This, of course, assumes that our preferred clustering does not contain predominantly singletons.

Kelley. Kelley et al.[17] describe a measure which they use for picking an optimal clustering of protein NMR ensembles. This method is also used by BCI's OPTCLUS program.[13] The Kelley measure balances the normalized "tightness" of the clusters at a particular level with the number of clusters (k_l) at that level. For a cluster level l, it is defined as

$$\mathrm{KELLEY}_l = (n-2)\left(\frac{\bar{d}_{wl} - \min(\bar{d}_w)}{\max(\bar{d}_w) - \min(\bar{d}_w)}\right) + 1 + k_l$$

where $\bar{d}_{wl}$ is the mean of distances between points in the same cluster at level l and $\min(\bar{d}_w)$ and $\max(\bar{d}_w)$ are the minimum and maximum of this value across all of the cluster levels, respectively. Singletons are excluded from the calculation, and the value of k_l is relied upon to sufficiently penalize cluster levels with large numbers of singletons. The smallest value of the Kelley measure should be chosen as the optimal level.

C-Index. Hubert and Levin[18] review a number of clustering level selection measures including the C-index. The C-index at a cluster level l is defined as

$$\mathrm{CINDEX}_l = \frac{\Gamma_l - \min(\Gamma)}{\max(\Gamma) - \min(\Gamma)}$$

where

$$\Gamma_l = {}^1/_2 \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} c_{ij}$$

n is the number of points, d_{ij} is the distance between point i and point j, and c_{ij} is a weighting factor applied to the distance. In the Hubert and Levin paper, this weighting factor is related to a concept of sequence in a psychological recall experiment. As implemented by Milligan et al.,[9] c_{ij} is simply 1.0 if points i and j lie within the same cluster or 0.0 if they are in different clusters. Γ is thus simply the sum of within-cluster distances across the set of clusters at a given level. It is this approach that we have adopted here. The max(Γ) and min(Γ) terms are the maximum and minimum values of Γ across all of the cluster levels, and therefore CINDEX simply records the normalized sum of within-cluster distances for each cluster level and is thus monotonic with the "tightness" term of the Kelley index. The minimum value of the C-index should be chosen for the optimum level.

Point Biseral. This index is described by Milligan.[15] It is defined for clustering l as

$$\text{POINTBIS}_l = (\bar{d}_{bl} - \bar{d}_{wl}) \left[\frac{n_{wl} n_{bl}}{n_d^2} \right]^{1/2} / \text{stddev}$$

where d_{bl} is the mean of the distances between points in different clusters at level l, d_{wl} is the mean of the distances between points in the same cluster at level l, n_{wl} is the number of distances between points in the same cluster, n_{bl} is the number of distances between points in different clusters, n_d is the total number of distances (i.e., $n(n-1)/2$), and stddev is the standard deviation of all distances from the mean. The level with the largest value is chosen as the optimal level.

Ball and Hall. This index, described by Milligan,[9] simply measures the mean distance between the centroid of each cluster and the points which belong to the cluster. A level is chosen which is the level immediately before (going from more clusters to fewer clusters) the level which has the biggest increase in mean distance over the previous level.

***W*/*B*.** The W/B measure is described by Milligan[15] who took it from a paper by McClain and Rao.[19] It is defined as

$$W/B_l = \bar{d}_{wl} / \bar{d}_{bl}$$

where $\bar{d}_{wl}$ and $\bar{d}_{bl}$ are as defined for the Point Biseral measure. The minimum value is taken as representing the optimal level.

Variance Ratio Criterion. Calinski and Harabasz[20] describe a measure known as the variance ratio criterion (VRC) which can be calculated for each level in a cluster hierarchy. The VRC measures the degree to which the mean between-point distances in clusters differ from the mean between-point distances in the whole set and is defined for a cluster level l as

$$\text{VRC}_l = \frac{\text{BGSS}_l / k_l - 1}{\text{WGSS}_l / n - k_l}$$

where k_l is the number of clusters at cluster level l and n is the number of points in the data set. BGSS_l is the between-group sum of squares, i.e., the sum of the distances between all of the points in different clusters at level l, and WGSS_l is the within-group sum of squares, i.e., the sum of the distances between all of the points which are in the same cluster at level l. For calculation, Calinski and Harabasz's alternative (but equivalent) definition is more useful:

$$\text{VRC}_l = \left[1 + \frac{n - k_l}{k_l - 1} a_l \right] \Big/ 1 - a_l$$

where a_l is the mean deviation of the within-group cluster distances from the mean of all of the squared distances, i.e.,

$$a_l = \frac{1}{\bar{d}^2 (n - k_l)} \sum_{i=1}^{k_l} (n_i - 1)(\bar{d}_2 - \bar{d}_i^2)$$

where n_i is the number of points in cluster i and $\bar{d}_i^2$ is the mean of the squared distances in cluster i.

Calinski and Harabasz suggest that the optimal cluster level will be the one where the VRC is at a local or absolute maximum, or at least where it undergoes a comparatively large increase. In the instance of there being several local maxima, they recommend choosing the one with the smallest value of k, i.e., the one with fewer clusters. This works with their test sets, which have a small number of points and a correspondingly small number of maxima, but in preliminary experiments we found that for our data sets this resulted in clusterings containing one very large and diverse cluster. Instead, we consider a local maximum to be a value where VRC drops off on either side. Out of the local maxima, we choose the one with the largest VRC value.

***γ*-Index.** This measure is described by Baker and Hubert[21] and has been used recently by Milligan.[15] It is defined as

$$\text{GAMMA}_l = \frac{s_l(+) - s_l(-)}{s_l(+) + s_l(-)}$$

where $s_l(+)$ represents the number of times where two points not clustered together at level l have a larger distance than two points which were clustered together, and $s_l(-)$ represents the number of times where they had a smaller distance. $s_l(+)$ and $s_l(-)$ may be calculated my making pairwise comparisons of distances in a distance matrix (where the x_y-th element is the distance between point x and point y). The larger the value of the γ-index, the better the clustering should be.

***τ* Index.** The τ index is described by Rohlf[22] and also by Milligan.[15] It is defined as

$$\text{TAU}_l = \frac{s_l(+) - s_l(-)}{[(n_d(n_d - 1)/2 - t_l)(n_d(n_d - 1)/2)]^{1/2}}$$

where $s_l(+)$ and $s_l(-)$ are as defined for the γ index, n_d is the number of distances in the distance matrix (i.e., $n(n-1)/2$, where n is the number of points in the data set), and t_l is the number of distance comparisons made where both distances are between points in the same cluster or both are between points in different clusters (i.e., the number of comparisons for which neither $s_l(+)$ nor $s_l(-)$ was incremented). The maximum value should be chosen as the optimal level.

***G*(+)-Index.** This index is described by Rohlf[22] and Milligan.[15] It is defined for clustering l as

$$G(+)_l = \frac{2s_l(-)}{n_d(n_d - 1)}$$

where $s_l(-)$ is as defined for the γ-index and n_d is the number of distances in the distance matrix (i.e., $n(n-1)/2$, where n is the number of points in the data set). The minimum value is taken as the optimal cluster level.

EXPERIMENTAL METHOD

For our experiments, we needed some way of establishing an "ideal" structural clustering of a data set, so that clusterings produced by the Ward's algorithm could be compared against this ideal to assess their goodness. Establishing the ideal clustering is not easy, because there is usually a level of subjectivity about which compounds should be clustered together. We created an ideal clustering of each of the seven data sets used by using our own chemical intuition to decide which structures are chemically related to one another. In some of the data sets used, this was a straightforward procedure, but in others (particularly the combinatorial sets), the decisions were much less easy to make, and other chemists may have made different choices. Still, we believe all of the clusterings represent a good and sensible partitioning of the structures.

Our experiments were performed using both data sets from our corporate compound collection and also compounds from the National Cancer Institute's (NCI's) anti-HIV data set, downloaded from the NCI's Internet site.[23] Our May 1997 version of this data set contained 32, 110 compounds which have been screened for anti-HIV1 activity in a cell-based assay,[24] including drugs, natural products, and dyes. The compounds are classified as active (100% inhibition of HIV-1 infection), moderately-active (at least 50% inhibition), and inactive (less than 50% inhibition). We chose the NCI data set because it is diverse (in that it contains a large variety of types of compound) but also contains distinct chemical series. In this regard it is similar to a typical pharmaceutical corporate database. Also, the nature of the screening results (active, moderately active, inactive) make it possible to extract compounds considered "actives" in the same way that one might from a high-volume screening run.

We extracted four sets of NCI compounds (NCI-A, NCI-B, NCI-C, and NCI-D) and three sets of proprietary compounds (PD-X, PD-Y, and PD-Z) to use for our experiments. Two of the data sets (data set NCI-A and NCI-B) contain a small number of compounds (55 and 79, respectively), specifically chosen to represent distinct series. The NCI-C, NCI-D, and PD-X data sets are larger and more diverse, while PD-Y and PD-Z are combinatorial sets and therefore have a significant element of commonality. For each data set, we manually identified the chemical series present and defined this as the ideal clustering of the compounds. A summary of the characteristics of the data sets together with the number of ideal clusters and singletons is given in Table 1.

To test the effectiveness of the level selection methods, we needed some way of assessing how close the clusterings represented by the levels chosen were to the ideal clustering. The traditional way of doing this is by using the Rand statistic.[25] In a particular clustering C_1, two points i and j may be either in the same cluster or in different clusters. If, in a second clustering C_2, points i and j are arranged in the same way (i.e., i and j are in the same cluster in C_1 and C_2, or they are in different clusters in both C_1 and C_2), then the two clusterings may be said to be consistent with respect to i and j. The value of Rand is defined as the fraction of pairs of points that are consistent between the two clusterings; that is

$$\mathrm{RAND}(C_1,C_2) = \frac{a+d}{a+b+c+d}$$

where a is the number of pairs of points that are clustered together in both clusterings and d is the number of pairs of points that are clustered into different clusters in both clusterings. b is the number of points that are paired together in the first clustering but not the second, and c is the number of points that are clustered together in the second clustering but not the first. A value of 1.0 represents a perfect match between the clusterings, and 0.0, a perfect mismatch. The Rand statistic has been shown to exhibit some undesirable properties both by our own experiments and in published work.[16] Particularly, it shows preference to clusterings containing a large number of clusters, especially with large data sets. This may be explained by the presence of the d term; the numerator and denominator become swamped by a very large value of d, so the measure becomes insensitive to changes in a, b, and c. An alternative measure is the Jaccard statistic, a simple modification to Rand which omits the d term from both numerator and denominator:

$$\mathrm{JACCARD}(C_1,C_2) = \frac{a}{a+b+c}$$

It is interesting to note that the Jaccard statistic is identical

Table 1. Characteristics of Each of the Data Sets

data set	no. of compds	desired no. of clusters	desired no. of singletons	data set source	comments
NCI-A	55	7	4	NCI anti-HIV data set	designed with clearly distinct but similar clusters
NCI-B	79	5	2	NCI anti-HIV data set	designed with clearly distinct but similar clusters
NCI-C	564	39	223	NCI anti-HIV data set	contains most of the active or moderately active compds from the set, clustering more subjective than for NCI-A or NCI−B
NCI-D	194	19	80	NCI anti-HIV data set	random subset of NCI−C.
PD-X	305	23	61	Parke-Davis compound library	diverse set of compds found active in a high throughput screen enzyme assay.
PD-Y	345	13	7	Parke-Davis combinatorial chemistry compounds	set of compounds derived from a single scaffold, developed for a single project
PD-Z	538	75	11	Parke-Davis combinatorial chemistry compounds	set of compds derived from a single scaffold, developed for a single project

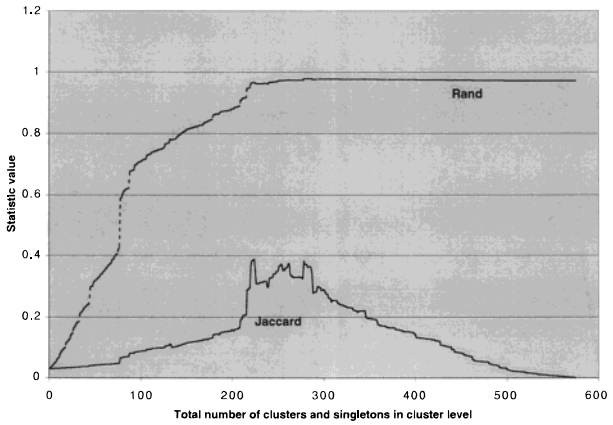

Figure 1. Behavior of Rand and Jaccard statistics with levels in the cluster hierarchy in a clustering of the NCI-C data set using Daylight fingerprints.

Table 2. Mean of Top Five Jaccard Statistics for Each Fingerprint Type and Each Data Set

	small, clear series		larger series, less well-defined			larger, combinatorial	
	NCI-A	NCI-B	NCI-C	NCI-D	PD-X	PD-Y	PD-Z
MACCS	0.837	*0.971*	0.470	*0.497*	0.392	0.168	0.186
BCI/D	0.806	0.963	0.360	0.422	*0.424*	0.204	0.327
Unity	0.817	0.940	*0.480*	0.437	0.376	0.262	0.256
Daylight	0.804	0.908	0.418	0.309	0.312	*0.472*	*0.555*
BCI/G	*0.839*	0.813	0.358	0.261	0.391	0.240	0.351

[a] Best scores for each data set are shown in italics.

to the Tanimoto coefficient,[2] a commonly used similarity measure in chemical information.

In his experiments, Milligan[15] found a high correlation between Jaccard and Rand (the method used for calculating correlation is not given by Milligan but is assumed to be a rank-correlation coefficient). However, Milligan used only small data sets (50 points) in which the *d* term does not become significant compared to the *a* value. We found a significant difference in behavior between two statistics when using them to evaluate cluster levels in a clustering of the 564-point NIC-C data set. Figure 1 shows how the behavior of the Rand, and Jaccard coefficients varies with the number of clusters in a cluster level. Jaccard clearly does not exhibit the strongly asymmetric behavior of the Rand statistic. Jaccard has been recommended by other studies,[16] so we chose to use it for this work.

RESULTS AND DISCUSSION

Relative Performance of Different Fingerprint Types. The Jaccard statistic was calculated between the ideal clustering and each level in the cluster hierarchy produced using different fingerprint types and data sets. Table 2 shows the mean of the top five values of the Jaccard statistic for each fingerprint method and data set. This both gives us an indication of the relative performance of the different fingerprint methods when used in conjunction with Ward's clustering (i.e., which fingerprints were able to produce cluster levels that most nearly represent the ideal clusterings, independently of which levels were selected by the level selection methods), and also gives us an upper bound for the performance of the cluster level selection methods. The data sets are classified according to their characteristics, as described in Table 1.

First, it is clear that all of the fingerprint types are able to produce clusterings similar to the ideal clustering, when applied to the NCI-A and NCI-B data sets, which are small and contain clear, separate series.

There is a greater divergence in performance when the clustering is applied to the larger data sets with less well-defined clusters (NCI-C, NCI-D, and PD-X). It appears that the methods that incorporate some kind of "structural key" type information that is independent of the data set used (MACCS, BCI/D, and to some extent Unity) generally outperform the fingerprints which are derived entirely from the structural paths present in the compounds in the data set (Daylight and BCI/G), although this distinction is only clearly shown in the NCI-D results. The fact that Unity fingerprints perform noticeably better than Daylight fingerprints is interesting, since the only major differences between them are that Unity partitions the fingerprint, and includes a small number of structural keys with frequency of occurrence information. The former difference may be significant, as the Euclidean distance calculation used in Ward's clustering becomes less sensitive with low bit densities, and the Unity fingerprint ensures that similar features are hashed into similar regions of the fingerprint.

Finally, it can be seen that the Daylight method produces significantly better clusterings than the other fingerprinting methods when applied to combinatorial data sets. MACCS keys perform especially poorly on these data sets, in contrast to their good performance with the other data sets. We may attribute this to the fact that MACCS has a small, fixed number of keys which may adequately characterize the differences between diverse compounds, but not discriminate the differences between closely related series. Daylight fingerprints contain information about all of the paths in the compounds and are therefore better suited to the discrimination of closely related compounds. It is unclear why Unity does not perform well on these data sets, since its fingerprints are very similar to Daylight, although it is possible that the hashing of similar features into similar areas of the fingerprint means that discrimination between similar features is lost when bit collisions occur. Similarly, one would expect the BCI/G method to perform well too. In this instance, we believe sparse distribution of bits in the fingerprint may be reducing its performance when used with Euclidean distance calculations.

Relative Performance of the Cluster Level Selection Methods. The Jaccard statistic was calculated between the ideal clustering and the top five levels selected by each cluster level selection method from the cluster hierarchy produced using different fingerprint types and data sets. We then gauged the success of each level selection method by dividing the mean Jaccard value of the selected levels with the upperbound values for the corresponding fingerprint type given in Table 2. Tables 3−5 shows these results expressed as a percentage for each fingerprint method and level selection method, when applied to each of the different types of data set (Table 3 shows the values for the small data sets of clear series, Table 4 shows the values for the larger data sets with less well-defined clusters, and Table 5 shows the values for the combinatorial sets). A value near 100% means that the level selection method chose levels very close to

Table 3. Percentage Success of the Level Selection Methods on the NCI-A and NCI-B Data Sets

fingerprint method	level selection method	NCI-A	NCI-B	mean	fingerprint method	level selection method	NCI-A	NCI-B	mean
MACCS	Point Biseral	100	100	100	Daylight	τ	41	68	54
Daylight	Point Biseral	100	88	94	Unity	τ	38	66	52
Unity	Point Biseral	100	87	94	MACCS	Kelley	68	34	51
BCI/G	Point Biseral	100	79	89	BCI/G	τ	39	58	48
BCI/D	Point Biseral	100	60	80	MACCS	τ	35	60	48
Daylight	Ball-Hall	68	87	78	BCI/D	τ	41	52	46
Unity	Ball-Hall	72	83	77	MACCS	VRC	57	34	46
BCI/G	Ball-Hall	61	90	75	Daylight	C-index	57	21	39
Daylight	γ	100	41	71	Daylight	*W*/*B*	57	21	39
Unity	γ	100	39	69	Daylight	*G*+	57	21	39
BCI/G	VRC	82	56	69	Unity	C-index	55	20	38
Unity	VRC	79	57	68	Unity	*W*/*B*	55	20	38
BCI/D	VRC	78	55	67	Unity	*G*+	55	20	38
BCI/D	Ball-Hall	48	83	66	BCI/D	C-index	53	20	36
BCI/D	γ	100	31	65	BCI/D	*W*/*B*	53	20	36
BCI/G	γ	100	24	62	BCI/D	*G*+	53	20	36
MACCS	Ball-Hall	60	64	62	BCI/G	C-index	50	19	35
Daylight	Kelley	80	38	59	BCI/G	*W*/*B*	50	19	35
MACCS	γ	100	18	59	BCI/G	*G*+	50	19	35
Daylight	VRC	77	40	58	MACCS	C-index	43	18	30
Unity	Kelley	81	32	57	MACCS	*W*/*B*	43	18	30
BCI/G	Kelley	74	39	56	MACCS	*G*+	43	18	30
BCI/D	Kelley	80	32	56					

Table 4. Percentage Success of the Level Selection Methods on the NCI-C, NCI-D, and PD-X Data Sets

fingerprint method	level selection method	NCI-C	NCI-D	PD-X	mean	fingerprint method	level selection method	NCI-C	NCI-D	PD-X	mean
MACCS	C-index	89	98	81	89	Unity	Point Biseral	31	25	72	42
MACCS	*W*/*B*	89	98	81	89	Unity	Kelley	31	22	72	41
BCI/D	C-index	96	97	72	88	Daylight	Kelley	31	34	52	39
BCI/D	*W*/*B*	96	97	72	88	Unity	Ball−Hall	48	27	42	39
MACCS	VRC	89	92	75	85	BCI/D	Ball−Hall	37	36	43	39
MACCS	τ	88	84	75	82	MACCS	Ball−Hall	43	37	36	38
MACCS	Kelley	84	71	69	75	Unity	τ	22	22	71	38
MACCS	Point Biseral	84	72	69	75	Daylight	Point Biseral	31	26	52	36
Unity	*W*/*B*	98	50	75	75	Daylight	τ	23	30	54	36
Daylight	C-index	83	58	81	74	Daylight	Ball−Hall	46	25	35	35
Daylight	*W*/*B*	83	58	81	74	Daylight	*G*+	21	58	24	34
Unity	C-index	98	48	75	74	BCI/G	τ	31	22	48	34
BCI/D	VRC	75	59	88	74	BCI/G	Ball−Hall	26	25	46	33
Unity	VRC	96	40	82	73	Unity	*G*+	11	48	33	31
BCI/G	C-index	95	44	80	73	BCI/G	Kelley	24	24	46	31
Daylight	VRC	86	31	84	67	BCI/G	*G*+	21	44	24	30
BCI/G	*W*/*B*	95	18	80	65	BCI/G	Point Biseral	24	17	46	29
MACCS	γ	38	98	36	57	BCI/D	γ	41	12	31	28
BCI/G	VRC	67	15	90	57	MACCS	*G*+	40	6	30	26
BCI/D	Kelley	45	26	94	55	Daylight	γ	23	18	34	25
BCI/D	*G*+	30	97	30	52	Unity	γ	13	10	38	20
BCI/D	Point Biseral	45	12	94	51	BCI/G	γ	23	17	17	19
BCI/D	τ	45	22	65	44						

the best level in the cluster hierarchy. In each case, the table is sorted by mean percentage value across the data sets.

Table 3 shows Point Biseral clearly outperforming the other methods for data sets with clearly-defined structural series, with the method perfectly selecting levels for the NCI-A data set and performing well on the NCI-B data set. The Ball−Hall method performs well on the NCI-B data set only, while the γ method performs well on the NCI-A set only. Table 4 shows that the C-index, *W*/*B*, VRC, and τ methods on MACCS and BCI/D fingerprints are the only ones to pick levels with over 80% overlap with the best level present. Table 5 shows the Kelley, Point Biseral, VRC, and τ methods to be the most effective when handling the combinatorial sets. Our results are thus showing that there is no one method that consistently performs better than the others, but that the most effective method is data set dependent.

If we are clustering data sets using a particular fingerprint type, we would like to be able to consistently use one level selection method no matter which data set we are clustering. Once we had shown that performance of the level selection methods is data set dependent, our next step was to consider the mean and worst-case performance of the methods across all of the data sets for each fingerprint type, to identify methods which had a reasonable mean performance, and which did not fail too badly in the worst case. Figure 2 shows a set of five graphs (a−e), one for each fingerprint method, with the mean and worst-case percentages shown. The *x*-axis is sorted so that the level selection methods with the best mean value are toward the left. In the case of the MACCS keys, the Point Biseral, and Kelley methods show good performance, with reasonable worst-case values. The BCI/D graph shows the VRC, Kelley, τ, and Ball−Hall methods maintain reasonable performance without the very poor worst-case performance of the other methods. In the Unity case, the VRC method stands out with a good mean performance and good worst-case value. The Daylight results show a small amount of variation in mean performance, but

Table 5. Percentage Success of the Level Selection Methods on the PD-Y and PD-Z Data Sets

fingerprint method	level selection method	PD-Y	PD-Z	mean	fingerprint method	level selection method	PD-Y	PD-Z	mean
BCI/G	Kelley	89	82	85	Unity	Ball-Hall	79	29	54
BCI/G	Point Biseral	89	82	85	BCI/D	Ball-Hall	84	23	54
MACCS	Kelley	81	87	84	Daylight	τ	47	56	52
BCI/G	VRC	71	90	80	Daylight	*G*+	63	42	52
Unity	τ	74	83	79	BCI/D	C-index	29	74	52
BCI/G	τ	91	61	76	BCI/D	*W/B*	29	74	52
Unity	Kelley	66	83	75	MACCS	γ	76	27	51
Unity	Point Biseral	66	83	75	Unity	*G*+	72	27	50
BCI/D	Kelley	53	97	75	Daylight	VRC	44	51	47
BCI/D	Point Biseral	53	97	75	Unity	VRC	38	53	46
Unity	γ	82	65	73	MACCS	VRC	30	61	45
BCI/D	*G*+	86	57	72	BCI/G	Ball−Hall	70	20	45
BCI/D	γ	81	61	71	Daylight	Ball−Hall	67	17	42
BCI/D	τ	43	97	70	MACCS	τ	37	45	41
MACCS	Point Biseral	38	87	63	MACCS	C-index	25	56	40
MACCS	Ball-Hall	81	43	62	MACCS	*W/B*	25	56	40
BCI/G	γ	79	43	61	BCI/D	VRC	30	50	40
MACCS	*G*+	82	39	60	Unity	C-index	26	50	38
Daylight	Kelley	48	70	59	Unity	*W/B*	26	50	38
Daylight	Point Biseral	48	70	59	Daylight	γ	67	6	36
BCI/G	C-index	41	75	58	Daylight	C-index	19	37	28
BCI/G	W/B	41	76	58	Daylight	*W/B*	19	37	28
BCI/G	*G*+	82	35	58					

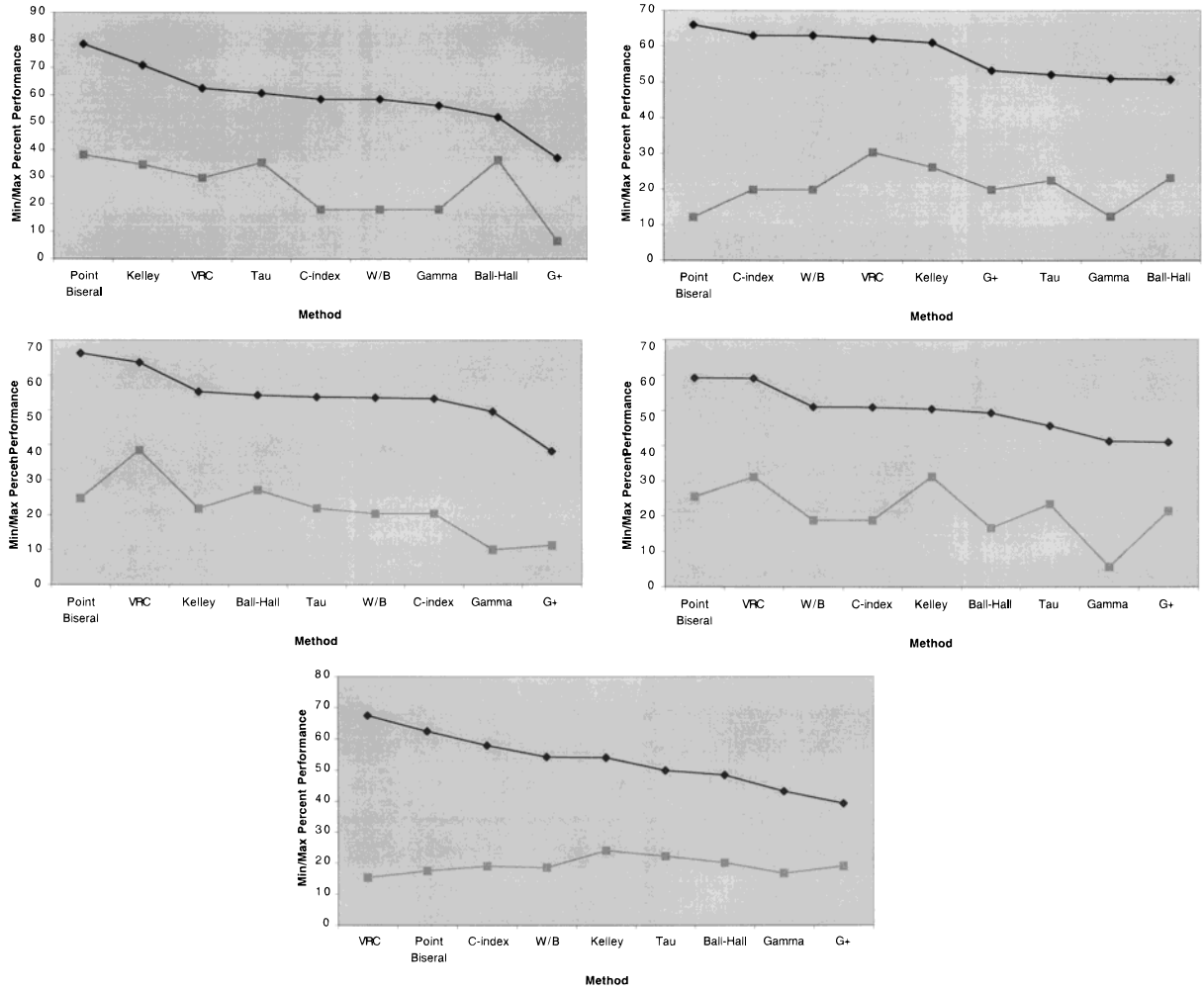

Figure 2. Graphs showing mean (black) and worst (grey) percent performances of each of the level selection methods for clusterings with (a, top left) MACCS keys, (b, top right) BCI/D fingerprints, (c, middle left) Unity fingerprints, (d, middle right) Daylight fingerprints, and (e, bottom) BCI/G fingerprints. The *x*-axes are sorted by decreasing mean performance.

considerable variation in worst-case performance, with the Calinksi and Kelley methods being the only ones with worst-case performance over 30%. Finally, the BCI/G results show the Kelley method probably has the best compromise between mean and worst-case performance, although both curves are fairly flat.

Table 6. Computational Complexity of Each Optimization Method

method	complexity
Kelley, C-Index, Ball−Hall	$<O(n^2)$
Point Biseral, VRC, *W*/*B*	$O(n^2)$
γ, τ, G(+)	$O(n^4)$

[a] n is the number of compounds being clustered, and c is the mean number of compounds in a cluster at a given level.

Another important factor in selecting a method is the computational complexity. Table 6 shows the complexities of the methods, and it is worthy of note that the performance is not related to the complexity of the method. Indeed, two of the methods with $O(n^4)$ complexity (γ, $G+$) are among the poorest methods.

CONCLUSIONS

Our experiments indicate that MACCS, Daylight, Unity, and BCI fingerprints all provide a good descriptor set for structural clustering when the compounds in the data set fall into clearly-defined structural classes. In data sets where the compounds are not so well grouped, there is some evidence that the methods that use dictionaries of predefined keys (MACCS, BCI/D, and to some extent Unity) are more effective than methods which generate set-dependent descriptors, although the difficulty in establishing what really is an ideal clustering in these sets makes us wary of making conclusions based on the small differences in the Jaccard similarities for these sets. However, experiments on the combinatorial sets clearly show Daylight fingerprints to be most effective. We may attribute this to the discriminating power of the fingerprints and perhaps to good hashing of bits about the fingerprint, although it is still unclear why it outperforms the Unity and BCI/G methods, and this deserves further investigation.

There is no one level selection method which consistently performs well across different types of data set. However, based on complexity, mean- and worst-case performance, we suggest that the following methods are likely to be most appropriate with each of the following fingerprint types: MACCS—Point Biseral or Kelley; BCI (default dictionary)—VRC, Kelley, or Point Biseral; BCI (generated dictionary)—Kelley; Unity—VRC; Daylight—VRC or Kelley. To date, we have only used the Kelley method in a practical setting where interactive performance is required and have found that it does a reasonably good job of identifying clusters relating to chemical series, although it does have a tendency to select a level containing a large cluster made up of several series and numerous smaller ones representing other series present. We intend to test some of the other methods that performed well in our experiments in everyday use. We are also interested in comparing the best (and selected) levels from Ward's clustering with the output of nonhierarchical clustering methods. This would help us judge whether it is realistic to expect a single level of a Ward's clustering to represent our ideal clustering and to what extent other algorithms may be able to produce better clusterings with the different fingerprint descriptors.

ACKNOWLEDGMENT

We would like to thank Alain Calvet and Christine Humblet for supporting this work, Paul Juneau for supplying the original Milligan and Cooper paper, and George Cowan, Michael Zhu, and Jeff Wu for discussions on the statistical aspects of this work.

REFERENCES AND NOTES

(1) 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.

(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press Ltd: Letchworth, U.K., 1987.

(3) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc*. **1963**, *58*, 236−244.

(4) Jarvis, R. A.; Patrick, E. A. Clustering using a Similarity Measure Based on Shared Near Neighbors. *IEEE Trans. Comput*. **1973**, *C-22*, 1025−1034.

(5) 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.

(6) Downs, G. M.; Willett, P. The Use of Similarity and Clustering Techniques for the Prediction of Molecular Properties. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devillers, J., Karcher, W., Eds.; ECSC: Brussels, 1991; pp 247−279.

(7) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and Representativity in Chemical Databases. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1−10.

(8) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(9) Milligan, G. W.; Cooper, M. C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **1985**, *50*, 159−179.

(10) Software and documentation available from Molecular Design Ltd., San Leandro, CA.

(11) Software and documentation available from Daylight Chemical Information Inc., Irvine, CA. E-mail: info@daylight.com.

(12) Software and documentation available from Tripos Associates, St Louis, MO. E-mail: support@tripos.com.

(13) Software and documentation available from Barnard Chemical Information Ltd., Sheffield, U.K. E-mail: barnard@bci1.demon.co.uk.

(14) Milligan, G. W. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* **1980**, *45*, 325−342.

(15) Milligan, G. W. A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. *Psychometrika* **1981**, *46*, 187−199.

(16) Milligan, G. W. Characteristics of Four External Criterion Measures. In *Proceedings of the 1982 NATO Advanced Studies Institute on Numerical Taxonomy*; Felsenstein, J. E., Ed.; Springer-Verlag: New York, 1983; pp 167−173.

(17) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally-Related Subfamilies. *Protein Eng*. **1996**, *9*, 1063−1065.

(18) Hubert, L. J.; Levin, J. R. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. *Psychol. Bull*. **1976**, *83*, 1072−1080.

(19) McClain, J. O.; Rao, V. R. CLUSTISZ: A program to test for the quality of clustering of a set of objects. *J. Mark. Res*. **1975**, *12*, 456−460.

(20) Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat*. **1974**, *3*, 1−27.

(21) Baker, F. B.; Hubert, L. J. Measuring the Power of Hierarchical Cluster Analysis. *J. Am. Stat. Assoc*. **1975**, *70*, 31−38.

(22) Rohlf, F. J. Methods of Comparing Classifications. *Annu. Rev. Ecol. Syst.* **1974**, *5*, 101−113.

(23) The data set may be downloaded from http://epnws1.ncifcrf.gov:2345/dis3d/aids_screen/aidspub.html.

(24) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J. P.; Shoemaker, R. H.; Boyd, M. R. New Soluble Formazan Assay for HIV-1 Cytopathic Effects: Application to High Flux Screening of Synthetic and Natural Products for AIDS Antiviral Activity. *J. Nat. Cancer Inst*. **1989**, *81*, 577−586.

(25) Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc*. **1971**, *66*, 846−850.

CI990086J