Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection

Aysha Al Khalifa, Maciej Haranczyk,[†] and John Holliday*

Department of Information Studies, University of Sheffield, Sheffield S1 4DP, U.K.

Received December 22, 2008

Several recent studies have compared the relative performance of a selection of similarity coefficients when applied to chemical databases represented by binary fingerprints. Considerable variation in performance, when used for (dis)similarity-based techniques, such as similarity searching, database clustering, and dissimilarity-based compound selection, has been reported, the reasons for which are closely related to molecular size. For many of these similarity coefficients, an alternative form can be derived which is applicable to sets of nonbinary data, such as calculated or measured physicochemical properties, or counts of substructural fragments. Here we report on several studies which have been undertaken to investigate the relative performance of twelve coefficients when applied to nonbinary data using such (dis)similarity-based techniques. Results suggest that no single coefficient is appropriate for all methodologies investigated and that the size bias detected with binary data is not as apparent when the data and, hence, coefficient are nonbinary in nature.

1. INTRODUCTION

Similarity-based methods have been employed in the chemical information industry in areas of similarity searching, property prediction, synthesis design, virtual screening, database analysis, and compound selection.[1] These techniques involve the comparison of characterizations of the respective molecules in order to quantify their structural similarity. The characterizations may be binary, such as a set of fingerprints in which the presence or absence of a set of substructural features contained within the compound are indicated in binary form with the values one indicating the presence of the features and the values zero indicating their absence; or they may be nonbinary, such as a set of physicochemical properties, measured chemical properties, or substructural feature counts. The degree of similarity between these characterizations, and hence it is intended, between the compounds themselves, is governed by a similarity coefficient, of which there are many forms.[2,3] These coefficients fall into three categories: association coefficients, commonly used on binary data and often normalized to lie with in a range zero to one; correlation coefficients, which measure the degree of correlation between the characterizations; and distance coefficients which quantify the degree of dissimilarity between characterizations. Here we use the term *similarity coefficient* when referring to all three categories.

For binary data, the most widely used coefficient is the Jacard/Tanimoto coefficient, but several alternatives have been investigated and compared in recent studies.[4−7] In particular, recent studies[4,5] have compared the relative performance of thirteen such coefficients which have been shown to exhibit complementary behavior when applied to similarity searches of chemical databases. As well as comparing their relative individual performance, further studies have investigated their use in combination using data fusion techniques.[8−10] More recently, a study was carried out by Haranczyk et al. to investigate the relative performance of these thirteen coefficients when applied to database clustering and dissimilarity-based compound selection methodologies.[11]

Many of the thirteen complementary coefficients have an alternative form which can be used to quantify the degree of similarity between nonbinary characterizations. In this study, we have extended the studies which have examined binary forms of the similarity coefficients by applying a set of nonbinary coefficients to nonbinary characterizations of a chemical structure database.

Similarity Search. The similar property principle[12] states that structurally similar compounds are likely to exhibit similar biological activity. As such, if an active compound is used to query a chemical structure database, it is expected that those which are more similar to the query compound are more likely to be similarly active. By ranking the results of similarity searches, enrichment has been found to increase significantly. Similarity searching has, as a result, been a fundamental concept in the drug discovery process for over a decade and improving its effectiveness continues to be an area of interest.

Previous studies[6,7,9,10] investigated whether enrichment could be improved by using alternative coefficients. The results of these studies showed that the variation in performance levels of these coefficients was mainly a result of the size, in terms of bit density, of the query itself, the database compounds that were members of the same active class, and the compounds that formed the rest of the database. Tests using binary data showed that, although the Tanimoto was generally the most effective coefficient, in terms of actives

* To whom correspondence should be addressed. E-mail: j.d.holliday@sheffield.ac.uk.

[†] Current address: Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

10.1021/ci8004644 CCC: $40.75 © 2009 American Chemical Society
Published on Web 04/30/2009

Table 1. Twelve Nonbinary Similarity Coefficients[a]

type	code in tables and figures	coefficient	formula	range
D	MM	Mean Manhattan	$(\sum\lvert x_{jk} - x_{jl}\rvert)/n$	∞ to 0
D	ME	Mean Euclidean	$[(\sum\lvert x_{jk} - x_{jl}\rvert^2)^{1/2}]/n$	∞ to 0
D	MSE	Mean Squared Euclidean	$(\sum\lvert x_{jk} - x_{jl}\rvert^2)/n$	∞ to 0
D	BC	Bray/Curtis	$(\sum\lvert x_{jk} - x_{jl}\rvert)/[\sum(x_{jk} + x_{jl})]$	1 to 0
A	Tan	Jaccard/Tanimoto	$(\sum(x_{jk}\cdot x_{jl}))/[\sum(x_{jk})^2 + \sum(x_{jl})^2 - \sum(x_{jk}\cdot x_{jl})]$	−1/3 to 1
A	Dic	Dice	$[2\sum(x_{jk}x_{jl})]/[\sum(x_{jk})^2 + \sum(x_{jl})^2]$	0 to 1
A	Fos	Fossum	$[n(\sum(x_{jk}x_{jl}) -(1/2))^2]/[\sum(x_{jk})^2\cdot\sum(x_{jl})^2]$	0 to ∞
A	SS1	Sokal/Sneath (1)	$[\sum(x_{jk}x_{jl})]/[2\sum(x_{jk})^2 + 2\sum(x_{jl})^2 - 3\sum(x_{jk}x_{jl})]$	0 to 1
A	Kul1	Kulczynski (1)	$[\sum(x_{jk}\cdot x_{jl})]/[\sum(x_{jk})^2 + \sum(x_{jl})^2 - 2\sum(x_{jk}x_{jl})]$	0 to ∞
A	Cos	Cosine/Ochiai	$[\sum(x_{jk}x_{jl}))/[(\sum(x_{jk})^2\cdot\sum(x_{jl})^2)^{1/2}]$	0 to 1
A	Sim	Simpson	$[\sum \min (x_{jk},x_{jl})]/[\min (\sum x_{jk}, \sum x_{jl})]$	0 to 1
C	Pea	Pearson	$[\sum(x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)]/[(\sum\overline{(x_{jk} - \bar{x}_k)^2\sum(x_{jl} - \bar{x}_l)^2)^{1/2}]}$	−1 to 1

[a] x_{jk} is the value of a descriptor in compound k at attribute j; x_{jl} is the value of a descriptor in compound l at attribute j; n is the total number of descriptors used for each compound; the limits of the summation are j from 1 to n. The coefficient types are: A-Association, D-Distance, and C-Correlation. Table based on Ellis et al. 1994.

retrieved, others were more applicable for certain size ranges. In particular, the Russell/Rao was highly effective when the query was large in size, and the Forbes more applicable for smaller queries. Furthermore, by combining selected coefficients, the search could be targeted toward a different size range relative to that of the query, in cases where it is necessary to alter the molecular bulk, for instance.

Clustering. Clustering is a classification methodology that groups objects together such that those objects within the same group or cluster are similar to each other but dissimilar to those in other clusters. When clustering is applied to a database of chemical structures, the clusters would be expected to define sets of structurally similar and, hence, because of the similar property principle, biologically similar compounds. Agglomerative clustering methods, as opposed to divisive methods that are not investigated here, fall into two main categories: hierarchical and nonhierarchical. Examples of hierarchical clustering, in which clusters start as singletons (i.e., each cluster contains one compound only) and are systematically merged together as they become more similar to each other, include Ward's method[13] and the group-average method.[14] An example of a nonhierarchical clustering methodology, in which cluster membership is determined by the degree of commonality between sets of nearest neighbors, is that of Jarvis−Patrick.[15]

A previous study[11] which compared the performance of the thirteen complementary coefficients using group-average hierarchical clustering and Jarvis−Patrick nonhierarchical clustering methods, based on the performance measure of eq 1, concluded that the correlation coefficients Pearson and Stiles, and the association coefficient Baroni−Urbani/Buser, appeared to outperform the others. In particular, the Russell/Rao, Forbes, and Simpson were found to be inferior for this methodology. (The Yule was found to be highly applicable for hierarchical clustering, but inferior for nonhierarchical.)

$$\frac{nA}{nC} \tag{1}$$

Wherein, for a given active class, nA is the number of active compounds in all active clusters (an active cluster being any cluster containing at least one member of the active class) and nC is the total number of compounds in the active clusters. A recent study[16] has described two alternative measures of clustering effectiveness, the first based on entropy, the distribution of actives across all clusters, and the second a probability-based measure which takes account of the number of inactive compounds in active clusters. For comparison with our earlier study, we continue to use the measure of eq 1.

Table 2. Eleven Bioactive Classes Indicating Their Class ID Code in the MDDR and the Number of Actives in the 20k Data Set

active class	class ID	code in tables and figures	number of actives
5HT3 antagonist	06233	5HT3	154
5HT1A agonist	06235	5HT1A	160
5HT reuptake inhibitor	06245	5HT_R	68
D2 antagonist	07701	D2	75
Renin inhibitor	31420	Renin	230
angiotensin II AT1 antagonist	31432	Angio	183
thrombin inhibitor	37110	Throm	168
substance P antagonist	42731	SubP	231
HIV-1 protesae inhibitor	71523	HIV-1	147
cyclooxygenase inhibitor	78331	Cox	130
protein kinase C inhibitor	78374	Kinase	83

Dissimilarity-Based Compound Selection. Dissimilarity-based compound selection (DBCS)[17] is aimed at identifying the subset of compounds from a chemical structure database that are the most structurally and, hence, biologically diverse. In drug discovery, this subset may be most appropriate for screening against a range of targets because screening the entire database would be too costly and time-consuming. The basic algorithm, shown in Figure 2, involves selecting an initial compound, randomly or systematically, and adding further compounds, one at a time, based on their maximal dissimilarity from those already selected.

A previous study[11] concluded that the size effects identified in similarity searching were exaggerated when the coefficients were applied to DBCS. In particular, those coefficients that were applicable to similarity searches in extreme size ranges, the Russell/Rao and Forbes, for example, were least appropriate for DBCS. Even the Tanimoto was shown to be inferior, a feature which has been reported and addressed by Fligner and co-workers.[18] The most appropriate measure for DBCS, based on the bit density distribution of selected compounds, appeared to be the Baroni−Urbani/Buser.[19]

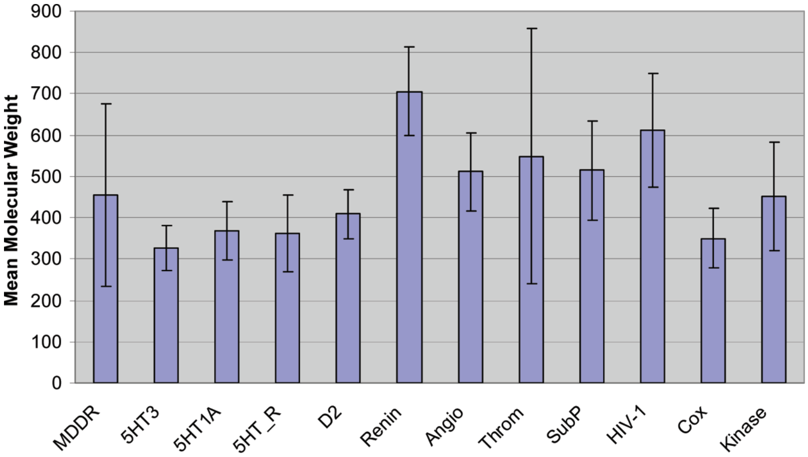

Figure 1. Mean molecular weight of 20K data set and active classes within it.

Table 3. Summary of Results for Similarity Searching[a]

Active class	Tan,Dic, Kul1,SS	Cos,Fos	Sim	Pea	M(S)E	BC	MM
Renin	***62.7***	52.7	25.6	**56.7**	**60.1**	**60.5**	**58.7**
HIV-1	***15.7***	10.3	**14.5**	10.8	**15.6**	***15.7***	**14.7**
Throm	***17.5***	13.8	15.1	**15.8**	**17.3**	**17.2**	**17.3**
SubP	9.2	7.1	***12.4***	7.7	8.7	9.1	8.6
Angio	17.1	12.1	***22.4***	13.1	17.1	7.3	17.3
Kinase	10.0	***15.1***	**13.7**	***15.1***	9.9	9.9	11.1
D2	**12.1**	10.0	4.0	10.7	**12.0**	**12.3**	***12.7***
5HT_R	6.6	6.2	***8.7***	5.6	**7.7**	6.8	**7.9**
5HT1A	6.5	**6.9**	4.4	**7.2**	**6.9**	**6.8**	***7.4***
Cox	**9.9**	**9.9**	5.9	9.1	***10.5***	***10.5***	***9.9***
5HT3	***22.9***	13.8	3.5	12.3	**22.6**	22.5	22.5
>90%	6	3	5	4	8	6	7

[a] Average percentage of actives retrieved in top 5%. The best performing coefficient for each class is shown in italic-boldface; those performing within 10% of this are shown in boldface.

2. NONBINARY SIMILARITY COEFFICIENTS

A review of similarity coefficients by Ellis et al.[3] discusses some 27 association, correlation and distance coefficients that can be applied to binary fingerprints and, in all but ten cases, have an equivalent nonbinary form. Of the 17 nonbinary coefficients, the Russell/Rao, Forbes, and Kulczynski(2) have been shown to produce values above and below the value of self-similarity and cannot, therefore, be ranked appropriately;[7] and the Mean Canberra and Divergence are not applicable to our data because they require positive data values. The resulting set of 12 coefficients is shown in Table 1; four of these are distance coefficients, one is a correlation coefficient, and seven are association coefficients.

3. METHODOLOGY

We used all twelve coefficients in similarity searching, clustering, and DBCS experiments to determine their relative performance when applied to these methodologies. In particular, we were interested in their relative effects across the different active classes, and whether these effects were as class-dependent as has been found in the case of binary descriptors. For clustering and DBCS, calculations were performed using a purpose-built set of programs to generate the required similarity matrices, one for each of the twelve similarity coefficients, and store these as binary files. Further programs were then used to carry out the required hierarchical and nonhierarchical clustering routines and DBCS.

1. Select a compound from the dataset at random or systematically and place it in the subset.
2. Identify the compound in the dataset that is most dissimilar to the compounds already in the subset.
3. Repeat Step 2 until the desired number of compounds is in the subset.

Figure 2. Dissimilarity-based compound selection routine.

Experimental Data. We used the same 20 000 compound subset from the MDL Drug Data Report (MDDR) database[20] used in the Haranczyk study as our data set. These compounds were then characterized by 378 structure-based property descriptors using the Molconn-Z package from eduSoft LC.[21] These descriptors include molecular connectivity indices, Kappa shape indices, E-State indices, standard topological indices, and subgraph counts such as paths and rings. We used a set of eleven bioactive classes (Table 2), previously reported by Hert et al.[22] and subsequently used by Haranczyk et al., to evaluate the similarity searching and clustering experiments. The bioactive classes represent a varied selection in terms of molecular size and structural homology, as indicated by the mean molecular weights shown in Figure 1.

Similarity Search Methodology. Ten compounds were selected randomly from each of the eleven active classes and these were used as queries in similarity searches of the 20K data set using all twelve similarity coefficients. The performance of each search was determined by the number of similarly active compounds retrieved in the top 5% of highest ranked data set compounds. Results were then compared across all active classes tested to determine whether there was any relationship between the best performing coefficients and the characteristics of the active class.

Clustering Methodology. For comparative purposes we used the same clustering methodologies used by Haranczyk et al.: group-average agglomerative clustering was therefore the hierarchical method used to cluster the 20K data set, and Jarvis−Patrick was the nonhierarchical method chosen. In group-average clustering, clusters are merged together based on the mean pairwise similarity, or average linkage, between members of the two clusters under investigation. In Jarvis−Patrick clustering, nearest neighbor lists of length *m* are derived for each item and items are placed in the same cluster if they are members of each others nearest neighbor list *and* if they have a minimum number of common items in their respective nearest neighbor lists. The clustering

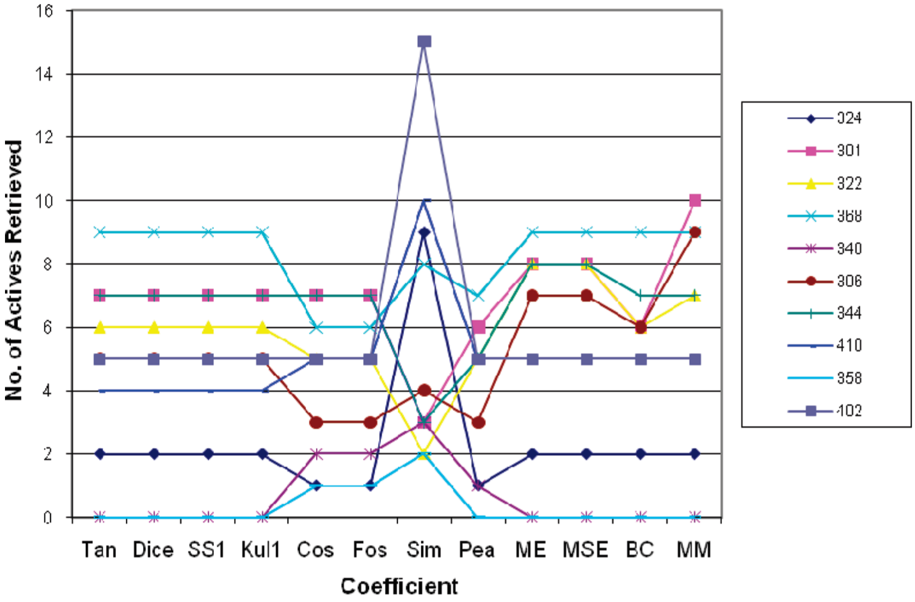

Figure 3. Actives retrieved in top 5% of data set for 5HT reuptake inhibitor class.

techniques were applied to the respective similarity matrices and, since the aim of clustering is to group together those compounds which are structurally similar, and hence similarly bioactive, the measure of performance given in eq 1 was used to indicate the effectiveness of each coefficient for each active class.

For hierarchical clustering, the performance measure was calculated throughout the clustering procedure at every hundredth iteration of the agglomeration process. This allows us to graphically compare coefficients across the whole clustering procedure for each active class, as indicated in the Results section.

Jarvis−Patrick clustering produces a varied number of clusters which, for a given similarity matrix, is dependent on the length of the nearest neighbor list (m) and the minimum number of common nearest neighbor items (p). Typical values for m and p are 14 and 8, respectively, but we varied these using values of 14, 16, 18, and 20 for m and 6 and 8 for p. The Jarvis−Patrick algorithm is known to produce many singleton clusters, clusters with only one member, which have the effect of increasing the performance measure. Results were therefore evaluated for all clusters, including singletons, and for all non-singletons. However, since the objective of clustering is to group similar objects together, it was understood that the non-singleton results were more valid.

Dissimilarity-Based Compound Selection Methodology. The basic DBCS algorithm of Figure 2 was applied to the respective similarity matrices to select a subset of structurally diverse compounds from the 20K data set. Here the initial compound chosen was that which was most dissimilar to the rest of the database. The best-case scenario would involve maximum class coverage with the minimum number of compounds in the subset. Since the full data set represents a total of 591 active classes, we decided to select 591 compounds from the data set with the ideal, though improbable, case being complete coverage of all classes. Indeed, since many compounds are active in more than one class, it is possible, though highly unlikely, to attain complete coverage with fewer selections than this. Our measure of performance was, then, the number of classes covered by the 591 compound subset. Again each of the twelve coefficients was used as the similarity metric.

4. RESULTS

Similarity Search Results. The rankings produced by the 1320 searches carried out indicate that several of the coefficients can be grouped due to their monotonic behavior. The Tanimoto, Dice, Kulczynski(1), and Sokal/Sneath(1) produce equivalent rankings, that is, they are monotonic, as do the Cosine and Fossum and, as would be expected, the Euclidean and Squared Euclidean.

Table 3 shows the relative performance of the coefficients in terms of the average percentage of actives retrieved for the ten searches in each active class/coefficient combination. (Those listed as monotonic are represented by a single column). Rows in the table, representing the active classes, have been ordered by mean molecular weight, in which the class with the largest mean molecular weight, Renin Inhibitors as illustrated in Figure 1, appears at the top. The best performing coefficient for each active class is shown in italic boldface and those performing within 10% of this are shown in boldface. The final row indicates those boldface (i.e., within 10% of the best performer) entries for each coefficient.

Several conclusions can be drawn from the results. First, that there is no single coefficient that consistently outperforms all others, although we can identify several poor performers such as the Cosine, Fossum, Pearson and, to some extent, the Simpson. Second, the size dependency, which is so obvious when binary data is applied to these searches, is not apparent in this case. The Tanimoto and Euclidean groups both appear to be equally effective for larger or smaller active classes. However, the Cosine, Fossum, and Simpson appear to be more effective for those classes whose mean molecular size is more similar to that of the data set.

The behavior of the Simpson is perhaps the most difficult to explain. Its performance is generally poor, but in a few individual cases, it noticeably outperforms all other coefficients. Figure 3 illustrates the results for ten individual

Table 4. Hierarchical Clustering Performance at (a) the 2000 Cluster Level, (b) the 1000 Cluster Level, and (c) the 500 Cluster Level[a]

active class	Tan	Dic	SS1	Kul1	Cos	Fos	Sim	Pea	ME	MSE	BC	MM
						(a) 2000 cluster level						
Renin	***0.2277***	***0.2277***	***0.2277***	**0.2160**	0.1420	0.1450	0.1410	0.0286	0.0128	0.0127	0.0128	0.0128
HIV-1	**0.0865**	0.0865	0.0865	0.0755	0.0486	0.0480	***0.0867***	0.0131	0.0081	0.0081	0.0082	0.0081
Throm	0.0662	0.0668	0.0662	0.0576	0.0406	0.0407	***0.1144***	0.0140	0.0093	0.0093	0.0094	0.0093
SubP	0.0619	0.0618	0.0619	0.0531	0.0402	0.0402	***0.1017***	0.0167	0.0128	0.0128	0.0129	0.0128
Angio	0.0654	0.0655	0.0653	0.0609	0.0423	0.0433	***0.0834***	0.0152	0.0102	0.0101	0.0102	0.0102
Kinase	**0.0635**	**0.0626**	***0.0640***	**0.0592**	0.0399	0.0405	0.0574	0.0057	0.0046	0.0046	0.0046	0.0046
D2	0.0420	0.0420	0.0420	0.0361	0.0248	0.0260	***0.0596***	0.0071	0.0042	0.0042	0.0042	0.0042
5HT_R	**0.0441**	**0.0441**	**0.0441**	**0.0430**	0.0385	0.0377	***0.0455***	0.0072	0.0038	0.0038	0.0038	0.0038
5HT1A	0.0438	0.0436	0.0438	0.0433	0.0378	0.0360	***0.0748***	0.0139	0.0089	0.0089	0.0089	0.0089
Cox	0.0494	0.0492	0.0494	0.0411	0.0403	0.0382	***0.0846***	0.0105	0.0072	0.0072	0.0073	0.0072
5HT3	0.0766	0.0766	0.0766	0.0663	0.0499	0.0516	***0.0892***	0.0162	0.0086	0.0086	0.0086	0.0086
>90%	4	4	4	4	0	0	9	0	0	0	0	0
						(b) 1000 cluster level						
Renin	***0.1474***	***0.1474***	***0.1474***	**0.1414**	0.0982	0.0983	0.0764	0.0235	0.0121	0.0121	0.0121	0.0121
HIV-1	***0.0419***	**0.0411**	***0.0419***	**0.0385**	0.0283	0.0283	**0.0393**	0.0120	0.0077	0.0077	0.0077	0.0077
Throm	0.0372	0.0371	0.0372	0.0328	0.0250	0.0263	***0.0511***	0.0118	0.0088	0.0089	0.0088	0.0088
SubP	**0.0381**	**0.0376**	**0.0381**	0.0335	0.0274	0.0282	***0.0410***	0.0142	0.0122	0.0122	0.0122	0.0122
Angio	**0.0404**	***0.0412***	**0.0404**	0.0357	0.0292	0.0310	**0.0401**	0.0133	0.0096	0.0096	0.0096	0.0096
Kinase	***0.0292***	**0.0287**	***0.0292***	0.0257	0.0191	0.0190	0.0241	0.0050	0.0044	0.0044	0.0044	0.0044
D2	0.0218	0.0221	0.0218	0.0190	0.0148	0.0147	***0.0252***	0.0058	0.0039	0.0040	0.0039	0.0039
5HT_R	***0.0256***	***0.0256***	***0.0256***	0.0228	0.0206	0.0220	**0.0233**	0.0051	0.0036	0.0036	0.0036	0.0036
5HT1A	0.0273	0.0268	0.0273	0.0264	0.0232	0.0228	***0.0360***	0.0117	0.0084	0.0084	0.0084	0.0084
Cox	0.0246	0.0245	0.0246	0.0216	0.0205	0.0225	***0.0324***	0.0091	0.0068	0.0069	0.0068	0.0068
5HT3	***0.0451***	***0.0451***	***0.0451***	0.0399	0.0304	0.0317	**0.0410**	0.0129	0.0081	0.0081	0.0081	0.0081
>90%	7	7	7	2	0	0	9	0	0	0	0	0
						(c) 500 cluster level						
Renin	**0.0988**	**0.0988**	***0.0992***	0.0892	0.0575	0.0671	0.0396	0.0187	0.0118	0.0118	0.0118	0.0118
HIV-1	***0.0242***	***0.0242***	**0.0242**	**0.0223**	0.0186	0.0169	0.0200	0.0089	0.0075	0.0075	0.0075	0.0075
Throm	**0.0227**	***0.0230***	**0.0227**	0.0207	0.0178	0.0186	**0.0222**	0.0103	0.0086	0.0086	0.0086	0.0086
SubP	***0.0256***	**0.0248**	***0.0256***	0.0216	0.0201	0.0203	**0.0237**	0.0129	0.0118	0.0119	0.0118	0.0118
Angio	**0.0255**	***0.0268***	**0.0255**	0.0253	0.0208	0.0209	0.0226	0.0125	0.0094	0.0094	0.0094	0.0094
Kinase	**0.0135**	***0.0136***	**0.0135**	0.0117	0.0101	0.0108	0.0102	0.0046	0.0043	0.0043	0.0043	0.0043
D2	**0.0122**	***0.0126***	**0.0122**	0.0117	0.0100	0.0102	**0.0124**	0.0048	0.0038	0.0039	0.0038	0.0038
5HT_R	**0.0047**	0.0042	***0.0048***	**0.0045**	0.0043	0.0042	**0.0044**	0.0038	0.0034	0.0034	0.0034	0.0034
5HT1A	**0.0177**	0.0174	0.0177	0.0163	0.0163	0.0151	***0.0192***	0.0107	0.0082	0.0082	0.0082	0.0082
Cox	**0.0161**	***0.0163***	**0.0161**	0.0150	0.0141	**0.0148**	0.0156	0.0087	0.0067	0.0067	0.0067	0.0067
5HT3	***0.0317***	**0.0316**	***0.0317***	0.0242	0.0200	0.0200	0.0227	0.0121	0.0079	0.0079	0.0079	0.0079
>90%	11	10	11	6	1	1	6	0	0	0	0	0

[a] The best performing coefficient for each class is shown in italic-boldface; those performing within 10% of this are shown in boldface.

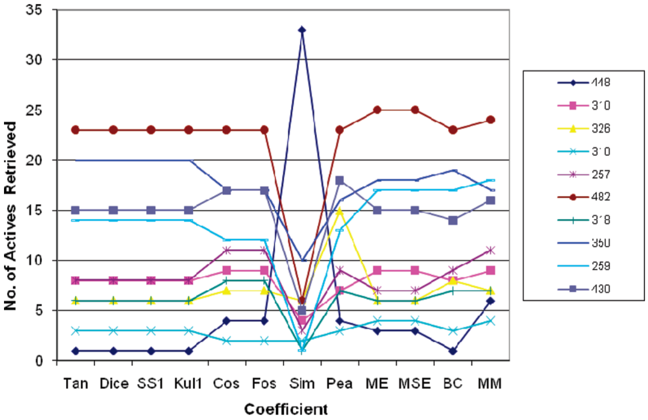

Figure 4. Actives retrieved in top 5% of data set for 5H1A agonist class.

searches, in terms of actives retrieved, for the 5HT reuptake inhibitor class. Here, the Simpson retrieves extremely well in several cases, increasing the mean retrieval rate considerably. In general, though, the relative performance is more like that illustrated in Figure 4, 5HT1A agonists, in which the single search is not enough to raise the mean retrieval. There does not appear to be a recognizable pattern for these individual high retrieval cases, either in terms of size or complexity of the query itself.

The size of each query molecule, in terms of molecular weight, is given in the legends of Figures 3 and 4. There is clearly no identifiable relationship between the size of the query and the performance of each coefficient; a relationship which has been clearly identified for binary coefficients. Equivalent conclusions were observed for all active classes tested.

Hierarchical Clustering Results. Table 4a−c shows the performance measures, (eq 1), for the 12 coefficients for each of the active classes. These are given for the 2000 cluster level, the 1000 cluster level, and the 500 cluster level. Three distinct groups of coefficient are identified in the hierarchical clustering stage, those that perform well, the Tanimoto, Dice, Sokal/Sneath, Kulczynski, and Simpson, those that perform poorly, the four distance coefficients and the Pearson, and those in between, the Cosine and Fossum. The results for the Tanimoto, Dice, and Sokal/Sneath are very similar, as would be expected since they are monotonic in similarity searches. Since the clustering method is value-based, the variation is more obvious. A similar relationship is found between the Cosine and Fossum.

Table 5. Nonhierarchical Clustering for Non-singleton Clusters[a]

active group	Tan	Dic	SS1	Kul1	Cos	Fos	Sim	Pea	ME	MSE	BC	MM
					(a) $m = 14$, $p = 8$							
number of NS clusters	251	251	250	250	892	897	580	1790	3	3	4	0
Renin	0.0328	0.0328	0.0326	0.0326	***0.0496***	**0.0494**	0.0123	0.0171	0.0115	0.0115	0.0115	
HIV-1	0.0122	0.0122	0.0122	0.0122	**0.0210**	**0.0211**	0.0078	0.0106	0.0073	0.0073	0.0073	
Throm	0.0160	0.0160	0.0159	0.0159	***0.0202***	***0.0202***	0.0090	0.0124	0.0084	0.0084	0.0084	
SubP	0.0161	0.0161	0.0160	0.0160	**0.0234**	***0.0235***	0.0124	0.0174	0.0116	0.0116	0.0116	
Angio	0.0154	0.0154	0.0154	0.0154	**0.0254**	***0.0255***	0.0098	0.0138	0.0092	0.0092	0.0092	
Kinase	0.0059	0.0059	0.0059	0.0059	***0.0109***	***0.0109***	0.0044	0.0064	0.0042	0.0042	0.0042	
D2	0.0057	0.0057	0.0057	0.0057	***0.0106***	***0.0106***	0.0040	0.0058	0.0038	0.0038	0.0038	
5HT_R	0.0051	0.0051	0.0051	0.0051	**0.0107**	***0.0110***	0.0036	0.0053	0.0034	0.0034	0.0034	
5HT1A	0.0112	0.0112	0.0112	0.0112	***0.0182***	***0.0182***	0.0086	0.0123	0.0080	0.0080	0.0080	
Cox	0.0110	0.0110	0.0110	0.0110	**0.0200**	***0.0202***	0.0070	0.0101	0.0065	0.0065	0.0065	
5HT3	0.0118	0.0118	0.0117	0.0117	***0.0226***	***0.0226***	0.0083	0.0120	0.0077	0.0077	0.0077	
>90%	0	0	0	0	11	11	0	0	0	0	0	0
					(b) $m = 20$, $p = 8$							
number of NS clusters	6	6	6	6	14	14	1078	1636	2	1	2	2
Renin	0.0115	0.0115	0.0115	0.0115	0.0115	0.0115	0.0135	***0.0162***	0.0115	0.0115	0.0115	0.0115
HIV-1	0.0073	0.0073	0.0073	0.0073	0.0073	0.0073	0.0086	***0.0103***	0.0073	0.0073	0.0073	0.0073
Throm	0.0084	0.0084	0.0084	0.0084	0.0084	0.0084	0.0099	***0.0118***	0.0084	0.0084	0.0084	0.0084
SubP	0.0116	0.0116	0.0116	0.0116	0.0116	0.0116	0.0136	***0.0162***	0.0116	0.0116	0.0116	0.0116
Angio	0.0092	0.0092	0.0092	0.0092	0.0092	0.0092	0.0108	***0.0127***	0.0092	0.0092	0.0092	0.0092
Kinase	0.0042	0.0042	0.0042	0.0042	0.0042	0.0042	0.0049	***0.0058***	0.0042	0.0042	0.0042	0.0042
D2	0.0038	0.0038	0.0038	0.0038	0.0038	0.0038	0.0044	***0.0053***	0.0038	0.0038	0.0038	0.0038
5HT_R	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0040	***0.0048***	0.0034	0.0034	0.0034	0.0034
5HT1A	0.0080	0.0080	0.0080	0.0080	0.0080	0.0080	0.0094	***0.0111***	0.0080	0.0080	0.0080	0.0080
Cox	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065	0.0076	***0.0090***	0.0065	0.0065	0.0065	0.0065
5HT3	0.0077	0.0077	0.0077	0.0077	0.0077	0.0077	0.0091	***0.0107***	0.0077	0.0077	0.0077	0.0077
>90%	0	0	0	0	0	0	0	11	0	0	0	0

[a] The best performing coefficient for each class is shown in italic-boldface; those performing within 10% of this are shown in boldface.

Table 6. Variation of Performance of Coefficients with Jarvis−Patrick Parameters Indicating the Number of Good Performances (Within 10% of Best Performer) Across All Classes[a]

m	*p*	Tan	Dic	SS1	Kul1	Cos	Fos	Sim	Pea	ME	MSE	BC	MM
							(a) all non-singleton clusters						
14	6	0 (22)	0 (22)	0 (22)	0 (22)	0 (37)	0 (36)	0 (835)	11 (1826)	0 (2)	0 (2)	0 (3)	0 (0)
14	8	0 (251)	0 (251)	0 (250)	0 (250)	11 (892)	11 (897)	0 (580)	0 (1790)	0 (3)	0 (3)	0 (4)	0 (0)
16	6	0 (10)	0 (10)	0 (10)	0 (10)	0 (23)	0 (23)	0 (988)	11 (1730)	0 (2)	0 (2)	0 (2)	0 (2)
16	8	0 (26)	0 (26)	0 (26)	0 (26)	0 (82)	0 (82)	0 (777)	11 (1750)	0 (3)	0 (3)	0 (2)	0 (2)
20	6	0 (3)	0 (3)	0 (3)	0 (3)	0 (9)	0 (9)	11 (1261)	11 (1557)	0 (1)	0 (1)	0 (2)	0 (2)
20	8	0 (6)	0 (6)	0 (6)	0 (6)	0 (14)	0 (14)	0 (1078)	11 (1636)	0 (2)	0 (1)	0 (2)	0 (2)
							(b) all clusters, including singletons						
14	6	0 (73)	0 (73)	0 (73)	0 (73)	0 (314)	0 (313)	11 (18731)	0 (12747)	10 (19964)	10 (19967)	10 (19963)	11 (20000)
14	8	0 (546)	0 (546)	0 (544)	0 (544)	0 (2576)	0 (2577)	11 (19238)	0 (14513)	11 (19978)	11 (19978)	10 (19964)	11 (20000)
16	6	0 (50)	0 (50)	0 (50)	0 (50)	0 (159)	0 (159)	10 (18298)	0 (11659)	10 (19953)	10 (19956)	10 (19952)	11 (19981)
16	8	0 (84)	0 (84)	0 (84)	0 (84)	0 (521)	0 (518)	11 (18872)	0 (12754)	11 (19968)	11 (19968)	10 (19955)	11 (19983)
20	6	0 (39)	0 (39)	0 (39)	0 (39)	0 (84)	0 (83)	6 (17373)	0 (9992)	9 (19929)	10 (19933)	10 (19941)	11 (19970)
20	8	0 (45)	0 (45)	0 (45)	0 (45)	0 (113)	0 (112)	11 (18064)	0 (10494)	10 (19941)	10 (19937)	10 (19941)	11 (19970)

[a] The number of clusters shown in parentheses.

The performance of the Simpson is, again, erratic. In the early stages of clustering, performance measures for the Simpson are high for nearly all classes, with the exception of the Renin Inhibitors. However, as clusters are agglomerated, and hence become larger, the performance drops significantly. This might be explained by the similarity search results in which the Simpson appears to identify a small number of highly similar actives. At an early clustering stage, these will cluster readily to produce a few highly active clusters. As the clusters become larger, the effect becomes reduced because of those actives whose Simpson similarity is low. With the Renin inhibitor class, there is a high degree of structural similarity and the high performance measures are reflected across all coefficients.

Figure 5 illustrates the comparative results for the angiotensin II AT1 antagonists, which are typical of most of the active classes. The last 2000 cluster levels are enlarged, showing the drop in performance of the Simpson coefficient. The high value in performance measure for several coefficients early on in the clustering strategy is caused by active singleton clusters. This effect is removed as clustering progresses, often producing a sharp decline in the measure.

Nonhierarchical Clustering Results. Table 5 a,b shows the performance measures for Jarvis−Patrick clustering using parameters $m = 14$, $p = 8$ and $m = 20$, $p = 8$, respectively, in which m is the length of the nearest neighbor list, p is the minimum number of common neighbors, and the singleton clusters have been omitted. Again, the best performing

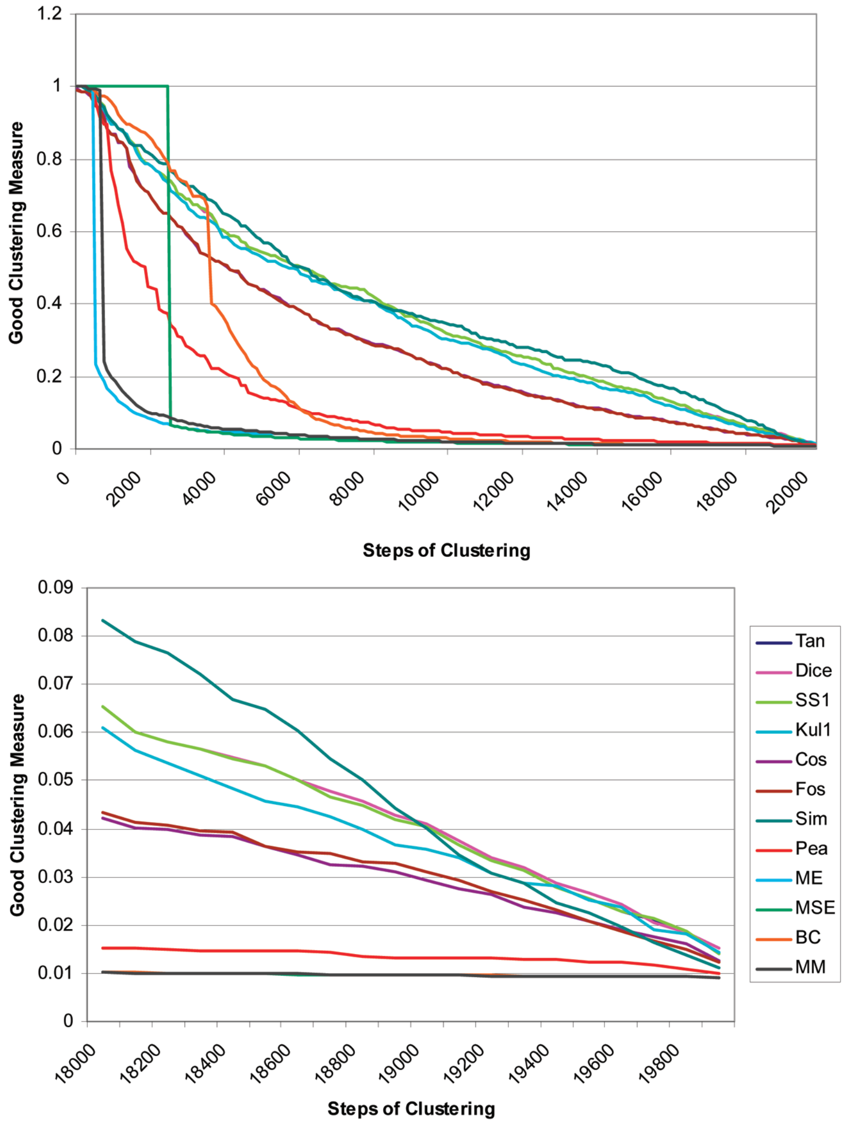

Figure 5. Hierarchical clustering of angiotensin II AT1 antagonists.

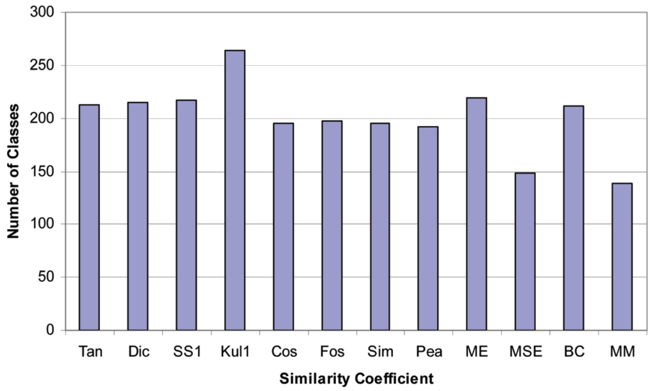

Figure 6. Classes identified using dissimilarity-based compound selection.

coefficient for each active class is indicated in italic-boldface, and those coefficients performing within 10% of this best score are shown in boldface. The final row indicates the number of times each coefficient appears as a boldface entry.

The first noticeable feature is the failure of the Mean Manhattan to cluster at all, with only singleton clusters being produced. Indeed, for all parameter combinations the distance coefficients clustered very poorly, the minimum number of clusters being 19920. When considering all clusters, including singletons, this tends to produce a high score using eq 1, since the active singleton clusters contribute significantly to the score.

Table 6a and b shows the number of high performing coefficients (i.e., those whose performance measure is within 10% of the best performer for each active class) for each parameter combination. The number of clusters is shown in parentheses, but it should be noted that the total number of clusters produced is always given in Table 6b; those shown in Table 6a are a subset of these. For non-singleton clusters, the Pearson coefficient appears to be the most effective; however, it does produce a considerable number of clusters. More realistic and meaningful clustering is performed with the standard parameters of $m = 14$, $p = 8$, using the Cosine or Fossum, giving 2576 and 2577 clusters respectively, over third of which are highly discriminating, non-singleton clusters.

The levels of performance compare well with those seen with binary data,[11] with similar figures being seen for our performance measure. The best performance values in Table

Table 7. Summary of four methods based on Tables 3−6 and Figure 6[a]

	Tan	Dic	SS1	Kul1	Cos	Fos	Sim	Pea	ME	MSE	BC	MM
similarity search	6	6	6	6	3	3	5	4	8	8	6	7
hierarchical clustering	22	21	22	12	1	1	24	0	0	0	0	0
nonhierarchical clustering	0	0	0	0	1.83	1.83	1.83	9.17	0	0	0	0
DBCS	213	215	217	264	195	198	195	192	219	148	212	139

[a] Hierarchical clustering figures taken as the sum of the three cluster levels, non-hierarchical clustering figures taken as the average of all results for non-singleton clusters.

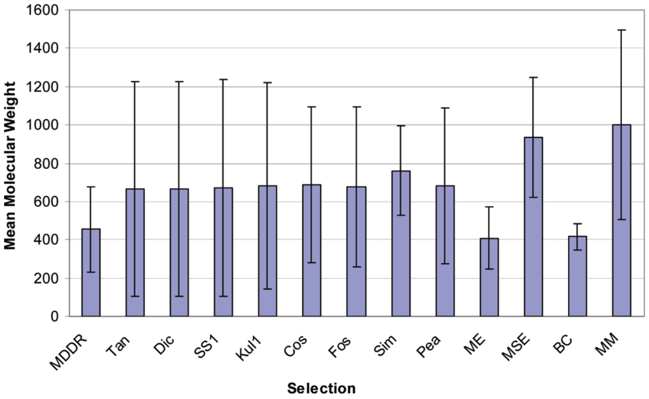

Figure 7. Mean molecular weights for compounds selected, with standard deviations.

5a (i.e., the italic entries) have a mean of 0.0212 (max, 0.0496; min, 0.0106), which compare to the equivalent best binary values which have a mean of 0.0180 (max, 0.0281; min, 0.0084), although this comparison does not relate to equivalent sets of coefficients. With binary data, however, the change in parameter values, particularly the length of the nearest neighbor list (*m*), did produce a more noticeable effect on the relative performance levels. Here, as would be expected, the main change is the decrease in the number of clusters with the increase in *m*.

Dissimilarity-Based Compound Selection Results. The aim of DBCS is to select a representative set of compounds that exhibits good structural and biological variety and also mimics the distribution of database characteristics. In terms of structural and biological variety, we have taken our ideal measure of performance as the complete selection of all 591 active class types using a set of 591 compounds. The results, shown in Figure 6, indicate that the Kulczynski(1) coefficient is the most effective with respect to biological diversity, with a coverage of 264 classes. Most other coefficients identify about 200 classes, with the exception of the Mean Squared Euclidean and the Mean Manhattan, which both performed poorly. Similar values were found when the binary version of DBCS[11] was investigated, with an average of around 200 classes being identified.

However, when considering the database characteristics, in this instance molecular weight, Figure 7 indicates that the Bray/Curtis and, in particular, the Mean Euclidean reflect the database distribution more realistically. Indeed all other coefficients tend to select larger compounds, many of which might not be considered as being drug-like without prior knowledge as their molecular weights exceed the upper limit identified in Lipinski's "rule of five".[23]

5. DISCUSSION AND CONCLUSIONS

A summary of the results for all four experimental methods, based on Tables 3−6 and Figure 6, is given in Table 7. This indicates the relative performance of each of the coefficients as, for similarity search, ME, MSE > MM > Tan, Dic, Kul1, SS1, BC > Sim > Pea > Cos, Fos, for a hierarchical clustering, Sim > Tan, SS1 > Dic > Kul1 > Cos, Fos > Pea, ME, MSE, BC, MM, for a nonhierarchical clustering (non-singleton clusters considered), Pea > Cos, Fos, Sim > Tan, Dic, SS1, Kul1, ME, MSE, BC, MM, and for DBCS, Kul1 > ME > SS1 > Dic > Tan > BC > Fos > Cos, Sim > Pea > MSE > MM.

Clearly, there is no single coefficient which works consistently well across all methodologies. It could be argued that the Cosine and Fossum, two coefficients found to exhibit near identical performance when applied to similarity searches using binary fingerprints,[4,9] would be more favorable for nonhierarchical clustering because of the number of clusters identified; or that the Mean Euclidean best represents the data set characteristics for DBCS. None of these are, however, consistent across all methodologies. Indeed, the four rankings above are shown to have no correlation at the 0.05 level of significance using the Kendall coefficient of concordance.

Perhaps the most versatile coefficients are the Tanimoto, Dice, Kulczynski(1), and Sokal/Sneath(1), which seem to perform comparably to others across all four methodologies. The binary forms of these four coefficients were again found to be near identical. The Simpson would appear to be the least robust because of the significant variation in performance, particularly, with regard to similarity searching, although it performs well in both clustering methods.

The most notable conclusion from these results is the lack of dependency on size of the molecules within the active classes. The size of the query (for similarity searching) and active class compounds, in terms of number of bits set or molecular weight, has a clear effect on the performance when binary coefficients are investigated.[11] This effect is not seen here, with comparable performance levels being observed across all active class sizes.

The experiments reported here complement those of a previous study which reported on the relative performance of these coefficients when applied to binary characterisations of the MDDR database. A more positive conclusion might be drawn, in future studies, by the use of several alternative and more appropriate databases or alternative characterisations.

ACKNOWLEDGMENT

We acknowledge the support from the Chemical Structure Association Trust for funding visits of Maciej Haranczyk to the University of Sheffield through the Jacques-Emile Dubois Grant. We also acknowledge Tripos Inc. and eduSoft, LC, for provision of the Molconn-Z software and MDL (now Symyx Technologies inc.) for provision of the MDDR database.

Note Added after ASAP Publication. This article was released ASAP on May 1, 2009 with minor errors in Tables 4 and 5. The correct version was posted on May 4, 2009.

REFERENCES AND NOTES

(1) Willett, P. Similarity methods in chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **2009**, *43*, 3–71.

(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(3) Ellis, D.; Furner-Hines, J.; Willett, P. Measuring the degree of similarity between objects in text retrieval systems. *Perspect. Inf. Manag.* **1994**, *3*, 128–149.

(4) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen.* **2002**, *5*, 155–166.

(5) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819–828.

(6) Holliday, J. D.; Salim, N.; Willett, P. On the magnitudes of coefficient values in the calculation of chemical similarity and dissimilarity. In *Chemometrics and Chemoinformatics*; ACS symposium series 894; Lavine, B. , Ed.; American Chemical Society: Washington, DC, 2005; pp 77−95.

(7) Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of similarity measures for searching the Dictionary of Natural Products database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449–457.

(8) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1–16.

(9) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435–442.

(10) Chen, J.; Holliday, J.; Bradshaw, J. A machine learning approach to weighting schemes in the data fusion of similarity coefficients. *J. Chem. Inf. Model.* **2009**, *49*, 185–194.

(11) Haranczyk, M.; Holliday, J. Comparison of similarity coefficients for clustering and compound selection. *J. Chem. Inf. Model.* **2008**, *48*, 498–508.

(12) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M. , Eds., John Wiley and Sons: New York, 1990.

(13) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

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

(15) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shard nearest neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034.

(16) Chu, C.-W.; Holliday, J. D.; Willett, P. Effect of standardization on chemical clustering and similarity searching. *J. Chem. Inf. Model.* **2009**, *49*, 155–161.

(17) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *78*, 65–84.

(18) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard−Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44*, 110–119.

(19) Baroni-Urbani, C.; Buser, M. W. Similarity of binary data. *Syst. Zool.* **1976**, *25*, 251–159.

(20) Symyx Technologies. MDL Drug Data Report. http://mdli.com/products/knowledge/drug_data_report (accessed 10 Nov 2008).

(21) *Molconn-Z*, version 4.0; eduSoft, LC: Ashland, VA, 2006.

(22) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(23) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

CI8004644