Comparison of Similarity Coefficients for Clustering and Compound Selection

Maciej Haranczyk† and John Holliday*,‡

Department of Chemistry, University of Gdańsk, 80-952 Gdańsk, Poland, and Department of Information Studies, University of Sheffield, Sheffield S1 4DP, U.K.

Received November 12, 2007

Recent studies into the use of a selection of similarity coefficients, when applied to searches of chemical databases represented by binary fingerprints, have shown considerable variation in their retrieval performance and in the sets of compounds being retrieved. The main factor influencing performance is the density distribution of the bitstrings for the active class, a feature which is closely related to molecular size. If this is the case when these coefficients are applied to similarity searches, then we would expect considerable variation in performance when applied to dissimilarity methods, namely clustering and compound selection. Here we report on several studies which have been undertaken to investigate the relative performance of 13 association and correlation coefficients, which have been shown to exhibit complementary performance in similarity searches, when applied to hierarchical and nonhierarchical clustering methods and to a compound selection methodology. Results suggest that the correlation coefficients perform consistently well for clustering and compound selection, as does the Baroni-Urbani/Buser association coefficient. Surprisingly, these often outperform the Tanimoto coefficient, while the Simple Match (effectively the complement of the Squared Euclidean Distance) performs very poorly.

INTRODUCTION

The similar property principle,[1] which states that structurally similar compounds are likely to exhibit similar biological activity, underlies many of the similarity techniques used to identify sets of active chemical compounds. When such techniques are applied to searches of compound databases using an active compound as the query, for example, the aim is to retrieve an increased proportion of similarly active compounds.[2,3] In order to carry out searches of this type, two components are required: a set of attributes or features which are used to uniquely characterize each compound and a measure which will quantify the degree of similarity between those sets of features. In the chemical information field the standard characterization takes the form of a bitstring, or binary fingerprint, in which the presence or absence of a set of chemical 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.

The degree of similarity between these binary fingerprints is measured using a similarity coefficient, of which there are many different forms, but the most widely used being the Tanimoto coefficient. The coefficients fall into three categories: association coefficients, commonly used with binary representations and often normalized to lie within the range zero (no common features) and unity (identical features); correlation coefficients, which measure the degree of correlation between the characterizations; and distance coefficients, which quantify the degree of dissimilarity between the characterizations and, when normalized, have the range zero (identity) and unity (no common features). Hereafter we shall refer to all three types as similarity coefficients, inferring that the complement of the distance coefficients can be regarded as a measure of structural similarity and vice versa.

In several recent studies,[4,5] which aimed to compare the relative performance of 13 of these similarity coefficients when applied to searches of chemical compound databases, it was found that there was considerable variation in the level of performance and in the sets of compounds being retrieved. The 13 coefficients used in these studies are shown in Table 1. In addition, subsequent studies have revealed a relationship between the bit density of the fingerprint and the performance of the similarity coefficients, where bit density is closely related to compound size.[6,7] In general, the Russell/Rao coefficient performs well in cases where the active compounds being retrieved are relatively large in size, while the Forbes works well when retrieving classes of relatively small compounds. If this level of variation is seen when different coefficients are applied to similarity searches, we would expect to observe likewise if they were applied to techniques which have a similarity component. The most extensively used techniques in this field which incorporate similarity comparisons are database clustering and dissimilarity-based compound selection. Indeed, the disparity between the sizes of compounds selected using the most common metrics for dissimilarity-based compound selection, the Tanimoto and the Euclidean Distance, has been observed.[8,9] The Tanimoto is known to select small compounds, while the Euclidean Distance generally selects large compounds, both of these being relative to the size of database compounds.

CLUSTERING AND COMPOUND SELECTION

Clustering is a classification methodology in which groups of similar objects are grouped together into clusters.[10] Objects

* Corresponding author e-mail: j.d.holliday@sheffield.ac.uk.
† University of Gdańsk.
‡ University of Sheffield.

10.1021/ci700413a CCC: $40.75
Published on Web 02/23/2008

Table 1. Thirteen Similarity Coefficients[a]

coefficient	formula	code used for tables and figures	coefficient type
Jaccard/Tanimoto	$\frac{a}{a+b+c}$	Tan	association
Russell/Rao	$\frac{a}{n}$	Rus	association
Simple Matching	$\frac{a+d}{n}$	SM	association
Baroni-Urbani/Buser	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	Bar	association
Ochiai/Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	Cos	association
Kulczynski(2)	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$	Ku2	association
Forbes	$\frac{n \times a}{(a+b)(a+c)}$	For	association
Fossum	$\frac{n\left(a-\frac{1}{2}\right)^2}{(a+b)(a+c)}$	Fos	association
Simpson	$\frac{a}{\min(a+b,a+c)}$	Sim	association
Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	Pea	correlation
Yule	$\frac{ad-bc}{ad+bc}$	Yul	correlation
Stiles	$\log_{10}\frac{n\left(\lvert ad-bc\rvert-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+c)(c+d)}$	Sti	correlation
Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	Den	correlation

[a] a is the number of bits common to both compounds, b and c are the number of bits unique to the query or database compound, respectively, d is the number of bits found in neither compound, and n is the fingerprint size ($n=a+b+c+d$).

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 1. Dissimilarity-based compound selection routine.

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	number of actives
5HT3 antagonist	06233	154
5HT1A agonist	06235	160
5HT reuptake inhibitor	06245	68
D2 antagonist	07701	75
renin inhibitor	31420	230
angiotensin II AT1 antagonist	31432	183
thrombin inhibitor	37110	168
substance P antagonist	42731	231
HIV-1 protesae inhibitor	71523	147
cyclooxygenase inhibitor	78331	130
protein kinase C inhibitor	78374	83

within a cluster are similar to each other but different from those from other clusters. When applied to databases of chemical structures, the clusters would then define sets of structurally similar and hence, if the similar property principle is obeyed, biologically similar compounds. There are several clustering methods, and these fall into two main categories: hierarchical and nonhierarchical. The most commonly used hierarchical clustering method is agglomerative hierarchical clustering in which the two most similar objects are merged into a single group, followed by the next two, and so on. Examples of agglomerative hierarchical clustering methods include Ward's method[11] and the group-average method.[10] This process is repeated, reducing the number of groups until the required number of groupings is achieved. Nonhierarchical methods, such as Jarvis-Patrick,[12] identify sets of nearest neighbors for each object and define the clusters by the degree of commonality between these sets.

Dissimilarity-based compound selection[13] seeks to identify a subset of the compound database which has the largest structural variation and hence contains many

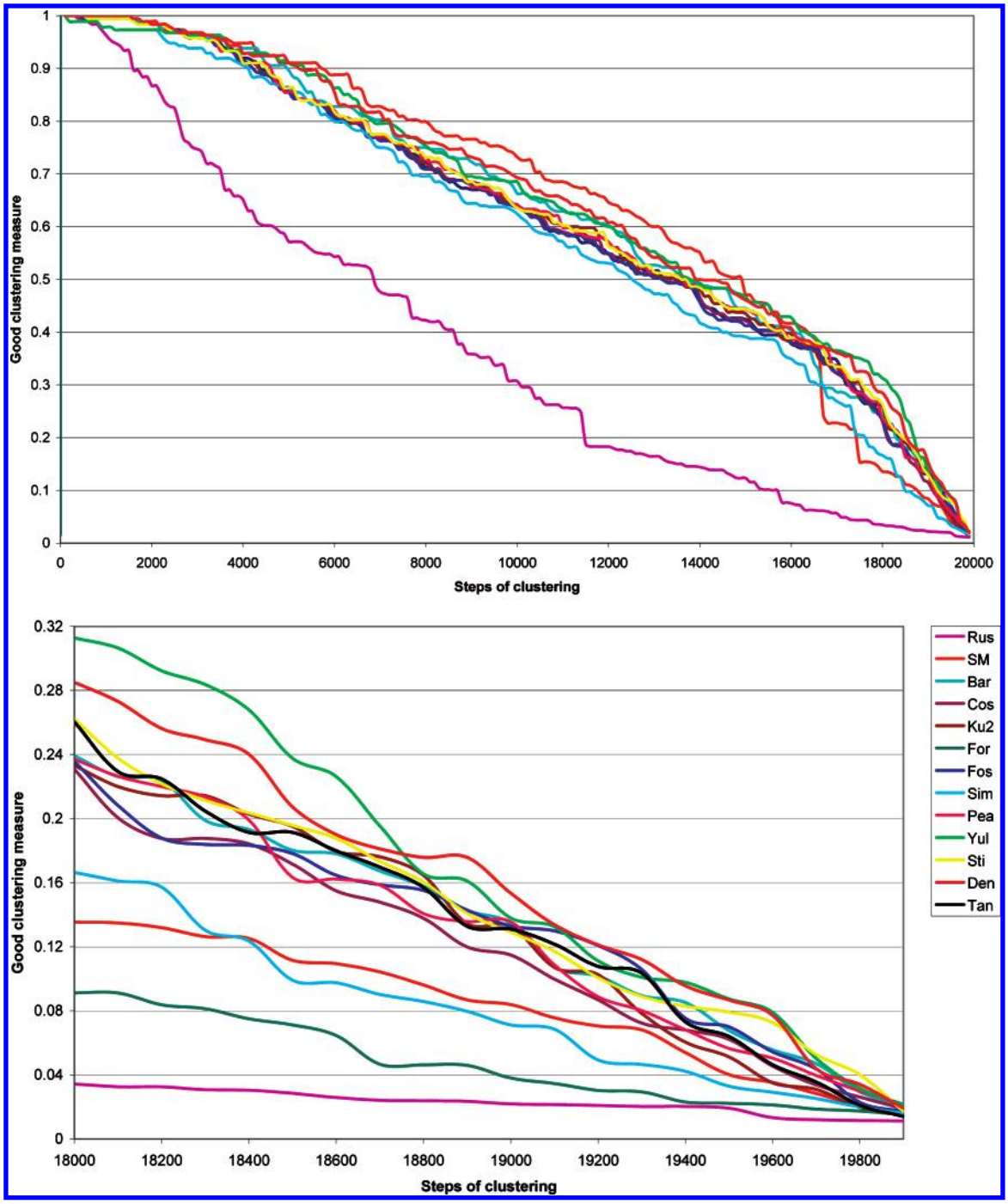

Figure 2. Hierarchical clustering of the angiotensin II AT1 antagonists (BCI1052).

different biologically active classes. In drug discovery, such sets may be appropriate for screening against a range of targets, an operation which would be too costly and time-consuming to apply to the whole database. The basic algorithm[14] involves selecting an initial compound, either randomly or by systematic means, and placing it in a subset. Further compounds are added to this subset one at a time, based on their maximal dissimilarity to those already selected.

EXPERIMENTAL DATA AND METHODOLOGY

Experiments reported here were carried out using a sample of 20 000 compounds from the MDL Drug Data Report (MDDR),[15] selected by taking every fifth compound in the database. These compounds were characterized by four fingerprint representations: the BCI standard 1052 fingerprints, available from Digital Chemistry;[16] the Daylight 1024 hashed fingerprints;[17] Unity standard 2D fingerprints from Tripos;[18] and Scitegic's level 4 extended connectivity fingerprints (ECFP4).[19] These representations were chosen as they include dictionary-based fingerprinting and hash-based fingerprint methodologies.

All calculations were performed using a purpose-built set of programs. Within this set, one program calculates the required similarity matrices and stores them on disk as binary files. These files are reused by two clustering programs: (i) one with so-called stored-matrix implementation of the sequential agglomerative hierarchical nonoverlapping (SAHN)

Table 3. Hierarchical Clustering Performance at a. the 2000 Cluster Level (BCI1052 Fingerprints), b. the 1000 Cluster Level (BCI1052 Fingerprints), and c. the 500 Cluster Level (BCI1052 Fingerprints)

active group	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
a. 2000 Cluster Level (BCI1052 Fingerprints)													
5HT3 antagonisis	0.1409	0.0256	0.0791	**0.1542**	**0.1522**	**0.1570**	0.0440	**0.1588**	0.1260	**0.1591**	***0.1628***	**0.1557**	**0.1528**
5HT1A agonists	0.1124	0.0253	0.0910	0.1174	0.1063	0.1121	0.0401	0.1154	0.0969	***0.1331***	0.1143	**0.1269**	**0.1292**
5HT reuptake inhibitor	**0.0798**	0.0141	0.0402	0.0640	0.0708	**0.0775**	0.0148	**0.0740**	0.0652	**0.0743**	***0.0808***	**0.0783**	0.0706
D2 antagonists	**0.0896**	0.0120	0.0554	**0.0859**	0.0810	**0.0879**	0.0178	0.0828	0.0690	0.0939	0.0748	**0.0866**	0.0727
renin inhibitor	0.1351	0.0416	***0.1580***	**0.1450**	0.1241	**0.1472**	0.1051	0.1267	0.1237	0.1285	**0.1570**	0.1335	0.1388
angiotensin II AT1 ant.	0.2603	0.0343	0.1355	0.2392	0.2308	0.2331	0.0913	0.2358	0.1664	0.2370	***0.3128***	0.2622	**0.2850**
thrombin inhibitor	0.1086	0.0272	0.1107	0.1296	0.1137	0.0996	0.0421	0.1033	0.0929	***0.1597***	0.1327	**0.1466**	0.1296
substance P antagonist	0.1151	0.0302	0.1088	***0.1486***	0.1196	0.1286	0.0480	0.1147	0.0925	**0.1359**	0.1314	0.1314	0.1197
HIV-1 protease inhibitor	0.0938	0.0212	0.0832	***0.1102***	0.0866	0.0963	0.0309	0.0894	0.0788	0.0877	0.0971	0.0831	0.0983
cyclooxygenase inhibitor	0.1285	0.0188	0.0573	0.1064	0.1186	***0.1432***	0.0236	0.1212	0.1059	0.1162	**0.1335**	0.1201	0.1105
protein kinase C inhibitor	0.0621	0.0134	0.0357	0.0652	**0.0732**	0.0534	0.0149	0.0716	0.0547	**0.0724**	**0.0744**	***0.0747***	0.0628
>90%	2	0	1	5	2	5	0	3	0	7	6	6	3
b. 1000 Cluster Level (BCI1052 Fingerprints)													
5HT3 antagonisis	0.0798	0.0162	0.0374	0.0795	0.0802	0.0711	0.0281	0.0809	0.0693	0.0783	0.0773	***0.0918***	0.0758
5HT1A agonists	0.0513	0.0132	0.0421	***0.0654***	0.0470	0.0586	0.0265	0.0540	0.0480	0.0535	**0.0604**	**0.0592**	0.0568
5HT reuptake inhibitor	0.0348	0.0081	0.0164	0.0288	0.0349	0.0331	0.0088	0.0350	0.0261	0.0360	0.0295	***0.0406***	0.0284
D2 antagonists	***0.0466***	0.0060	0.0224	0.0385	0.0324	0.0406	0.0127	0.0307	0.0338	0.0388	0.0415	0.0404	0.0379
renin inhibitor	0.0985	0.0268	0.0915	**0.1108**	0.0952	***0.1125***	0.0496	0.0920	0.0839	0.0941	0.0958	0.0978	**0.1072**
angiotensin II AT1 ant.	0.1310	0.0219	0.0839	0.1347	0.1147	0.1328	0.0380	0.1329	0.0712	0.1356	0.1379	0.1291	***0.1534***
thrombin inhibitor	***0.0691***	0.0188	0.0518	0.0604	0.0570	**0.0654**	0.0246	**0.0664**	0.0499	0.0585	**0.0642**	**0.0688**	**0.0659**
substance P antagonist	0.0643	0.0226	0.0559	***0.0726***	0.0609	0.0640	0.0280	0.0614	0.0466	0.0644	0.0611	**0.0698**	0.0643
HIV-1 protease inhibitor	**0.0581**	0.0138	0.0508	***0.0643***	**0.0599**	**0.0604**	0.0196	0.0533	0.0476	**0.0617**	**0.0596**	0.0544	**0.0604**
cyclooxygenase inhibitor	***0.0591***	0.0103	0.0227	0.0505	0.0446	**0.0575**	0.0154	0.0495	0.0516	**0.0539**	**0.0582**	**0.0564**	**0.0552**
protein kinase C inhibitor	**0.0372**	0.0074	0.0159	0.0318	0.0314	0.0280	0.0097	**0.0375**	0.0286	0.0307	0.0309	**0.0351**	***0.0379***
>90%	5	0	0	4	1	4	0	2	0	2	4	7	6
c. 500 Cluster Level (BCI1052 Fingerprints)													
5HT3 antagonisis	0.0365	0.0133	0.0240	***0.0437***	0.0359	**0.0394**	0.0228	0.0382	0.0363	0.0356	**0.0418**	**0.0431**	0.0385
5HT1A agonists	0.0244	0.0114	0.0264	**0.0383**	0.0286	0.0327	0.0232	0.0327	0.0293	0.0321	0.0281	**0.0365**	***0.0400***
5HT reuptake inhibitor	0.0125	0.0059	0.0111	***0.0200***	0.0162	0.0148	0.0072	**0.0184**	0.0120	0.0171	0.0173	**0.0194**	0.0173
D2 antagonists	***0.0216***	0.0051	0.0117	**0.0194**	0.0150	0.0161	0.0093	0.0179	0.0179	**0.0194**	**0.0199**	0.0184	**0.0194**
renin inhibitor	0.0558	0.0218	0.0716	***0.0898***	0.0652	0.0573	0.0342	0.0633	0.0533	0.0777	0.0788	0.0718	0.0799
angiotensin II AT1 ant.	0.0640	0.0192	0.0402	0.0676	0.0622	0.0512	0.0222	0.0700	0.0329	0.0566	***0.0874***	**0.0792**	**0.0867**
thrombin inhibitor	0.0333	0.0127	0.0285	**0.0352**	0.0312	0.0344	0.0179	**0.0381**	0.0324	**0.0376**	***0.0389***	**0.0365**	**0.0366**
substance P antagonist	0.0364	0.0160	0.0304	***0.0422***	0.0363	0.0331	0.0225	**0.0400**	0.0267	0.0349	**0.0390**	0.0352	0.0376
HIV-1 protease inhibitor	0.0280	0.0091	0.0261	***0.0465***	0.0357	0.0280	0.0131	0.0325	0.0320	0.0393	0.0415	0.0350	0.0360
cyclooxygenase inhibitor	**0.0270**	0.0087	0.0145	**0.0274**	**0.0274**	0.0252	0.0123	0.0255	0.0241	0.0248	***0.0291***	**0.0285**	**0.0277**
protein kinase C inhibitor	0.0143	0.0053	0.0110	***0.0193***	0.0162	0.0148	0.0079	0.0135	0.0149	0.0133	**0.0177**	**0.0173**	**0.0183**
>90%	2	0	0	10	1	1	0	3	0	2	7	7	6

Table 4. Summary of Relative Coefficient Performance for Three Cluster Levels: 2000, 1000, and 500 Clusters

fingerprint type	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
2000 Cluster Level													
BCI	2	0	1	5	2	5	0	3	0	7	6	6	3
Daylight	5	0	2	4	3	6	1	4	2	7	5	3	4
Unity	3	0	0	4	9	3	0	2	1	4	1	4	1
EFP4	5	0	2	4	3	6	1	4	2	7	5	3	4
1000 Cluster Level													
BCI	5	0	0	4	1	4	0	2	0	2	4	7	6
Daylight	2	0	2	3	4	1	1	2	0	4	3	1	2
Unity	3	0	1	4	4	4	0	3	0	8	2	8	1
ECFP4	2	0	2	3	4	1	1	2	0	4	3	1	2
500 Cluster Level													
BCI	2	0	0	10	1	1	0	3	0	2	7	7	6
Daylight	0	0	1	3	2	2	1	3	0	5	3	3	2
Unity	4	0	0	5	3	2	0	5	1	5	2	8	3
ECFP4	0	0	1	3	2	2	1	3	0	5	3	3	2
Sum	33	0	12	52	38	37	6	36	6	60	44	54	36

clustering method[20] and (ii) the other with the implementation of the Jarvis-Patrick clustering method described below; and by a compound selection program.

CLUSTERING

The group-average agglomerative approach was the method chosen for hierarchical clustering. In this method, clusters are merged together based on the mean pairwise similarity between members of the two clusters under consideration. For nonhierarchical clustering, the Jarvis-Patrick method was chosen. In Jarvis-Patrick clustering, nearest neighbor lists of length m are derived for each compound, using similarity techniques, and two compounds are placed in the same cluster if they are a member of each other's nearest neighbor list and if they have at least p neighbors common to both lists (where $p<m$). Typical values for p and m are 8 and 14, respectively, but several values were used in our tests in order to reduce the number of singleton clusters (clusters containing just one object), a feature which is common in Jarvis-Patrick clustering.

Table 5. Nonhierarchical Clustering at a. $m=14$, $p=8$ (BCI1052 Fingerprints) and b. $m=20$, $p=8$ (BCI1052 Fingerprints)

active group	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
a. $m=14$, $p=8$ (BCI1052 Fingerprints)													
number of clusters	2095	2235	2211	2097	2079	2057	2616	2088	2456	2103	2273	2101	2092
5HT3 antagonisis	**0.0182**	0.0139	***0.0188***	**0.0186**	**0.0182**	**0.0181**	0.0132	**0.0182**	0.013	**0.0184**	0.0164	**0.0184**	**0.0186**
5HT1A agonists	**0.0190**	0.0145	***0.0194***	**0.0192**	**0.0190**	**0.0188**	0.0136	**0.0190**	0.0135	**0.0191**	0.0171	**0.0191**	**0.0192**
5HT reuptake inhibitor	**0.0082**	0.0062	***0.0084***	**0.0083**	**0.0082**	**0.0081**	0.0058	**0.0082**	0.0057	**0.0083**	0.0073	**0.0083**	**0.0083**
D2 antagonists	**0.0089**	0.0068	***0.0092***	**0.0090**	**0.0089**	**0.0089**	0.0065	**0.0089**	0.0063	**0.0090**	0.0080	**0.0090**	**0.0091**
renin inhibitor	**0.0272**	0.0205	***0.0279***	**0.0276**	**0.0272**	**0.0270**	0.0198	**0.0272**	0.0194	**0.0275**	0.0246	**0.0274**	**0.0277**
angiotensin II AT1 ant.	**0.0215**	0.0163	***0.0222***	**0.0219**	**0.0215**	**0.0213**	0.0157	**0.0215**	0.0153	**0.0217**	0.0193	**0.0217**	**0.0219**
thrombin inhibitor	**0.0198**	0.0151	***0.0204***	**0.0201**	**0.0198**	**0.0197**	0.0144	**0.0198**	0.0142	**0.0199**	0.0180	**0.0199**	**0.0201**
substance P antagonist	**0.0273**	0.0207	***0.0281***	**0.0277**	**0.0273**	**0.0272**	0.0198	**0.0273**	0.0194	**0.0275**	0.0246	**0.0275**	**0.0276**
HIV-1 protease inhibitor	**0.0176**	0.0132	***0.0181***	**0.0179**	**0.0176**	**0.0174**	0.0126	**0.0176**	0.0124	**0.0177**	0.0158	**0.0177**	**0.0179**
cyclooxygenase inhibitor	**0.0154**	0.0117	***0.0158***	**0.0156**	**0.0154**	**0.0153**	0.0112	**0.0154**	0.0110	**0.0156**	0.0138	**0.0156**	**0.0155**
protein kinase C inhibitor	**0.0099**	0.0075	***0.0102***	**0.0101**	**0.0099**	**0.0098**	0.0070	**0.0099**	0.0070	**0.0100**	0.0090	**0.0100**	**0.0101**
>90%	11	0	11	11	11	11	0	11	0	11	0	11	11
b. $m=20$, $p=8$ (BCI1052 Fingerprints)													
number of clusters	1124	2243	1175	1106	1141	1164	2462	1153	2399	1133	1469	1121	1124
5HT3 antagonisis	0.0172	**0.0196**	0.0128	0.0124	0.0169	0.0162	0.0144	0.0174	***0.0208***	0.0145	0.0172	0.0148	0.0136
5HT1A agonists	0.0182	**0.0196**	0.0133	0.0128	0.0171	0.0162	0.0149	0.0176	***0.0218***	0.0145	0.0185	0.0146	0.0133
5HT reuptake inhibitor	0.0081	**0.0084**	0.0058	0.0055	0.0074	0.0072	0.0064	0.0077	***0.0094***	0.0076	0.0077	0.0065	0.0060
D2 antagonists	0.0083	**0.0095**	0.0063	0.0061	0.0079	0.0076	0.0071	0.0081	***0.0103***	0.0071	0.0083	0.0072	0.0063
renin inhibitor	0.0230	0.0265	0.0191	0.0187	0.0216	0.0211	**0.0285**	0.0216	***0.0307***	0.0223	0.0260	0.0193	0.0190
angiotensin II AT1 ant.	0.0192	0.0225	0.0156	0.0142	0.0177	0.0172	0.0170	0.0181	0.0246	0.0201	***0.0308***	0.0172	0.0163
thrombin inhibitor	0.0164	0.0200	0.0138	0.0130	0.0155	0.0157	0.0158	0.0159	***0.0230***	0.0143	0.0185	0.0141	0.0141
substance P antagonist	0.0218	0.0272	0.0186	0.0187	0.0202	0.0200	0.0216	0.0202	***0.0309***	0.0219	0.0268	0.0190	0.0186
HIV-1 protease inhibitor	0.0145	0.0176	0.0122	0.0119	0.0134	0.0137	0.0139	0.0137	***0.0203***	0.0126	0.0170	0.0124	0.0127
cyclooxygenase inhibitor	0.0115	**0.0161**	0.0106	0.0099	0.0118	0.0132	0.0121	0.0117	***0.0173***	0.0117	0.0147	0.0118	0.0105
protein kinase C inhibitor	0.0077	**0.0106**	0.0068	0.0062	0.0074	0.0077	0.0078	0.0075	***0.0113***	0.0070	0.0089	0.0070	0.0069
>90%	0	6	0	0	0	0	1	0	10	0	1	0	0

The two clustering techniques were applied using each of the 13 coefficients of Table 1 as the similarity metric and using each of the four fingerprint representations, giving a total of 52 clustering procedures for each technique.

As mentioned, the aim of a clustering process is to group together compounds which are structurally similar and should, therefore, be similarly bioactive. A measure of performance is then indicated by the efficiency in grouping compounds which exhibit a known biological activity. In order to evaluate performance of the clusters which have been produced, we require a bioactive class and a performance measure. We have chosen not just one but a selection of eleven bioactive classes, previously reported by Hert et al.,[21] to carry out our evaluation; these are shown in Table 2. It is expected that results will vary from class to class, and this may well be dependent on the coefficient being applied, as has been found to be the case in similarity searching. The performance measure we have chosen is as follows

$$\frac{nA}{nC}$$

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. The ideal clustering situation would find all actives as exclusive members of one, or more, cluster, giving a performance measure of unity. The worst case would see active compounds distributed evenly across all clusters, giving a measure approaching *nA*/20K (for *nA* clusters), in the present study.

HIERARCHICAL CLUSTERING

Due to the nature of hierarchical clustering, the cluster level at which to evaluate the methodology is, initially, unknown. In a standard procedure of this type, either a decision would be made about how many clusters are required or a stopping rule, such as that of Mojena[22] which uses statistical techniques to identify natural clusters, would be applied. In our experiments, we have applied our performance measure throughout the clustering procedure, taking measurements at every hundredth iteration of the agglomeration process. This has enabled us to compare coefficients across the entire procedure as shown in the graphs and tables of the Results section.

NONHIERARCHICAL CLUSTERING

Nonhierarchical clustering was carried out using the Jarvis-Patrick clustering algorithm. The number of clusters produced varies from coefficient to coefficient and also with changes to the parameter values *m* and *p*. The clustering was repeated using the nearest neighbor lists of length 14, 16, and 20 and common nearest neighbor values of 6 and 8. The Jarvis-Patrick algorithm is known to have the drawback of producing many singletons, and these would clearly increase the performance measure above as only active singletons would contribute to the score. Results were evaluated for all clusters, including singletons, and for all nonsingleton clusters. However, since the objective of clustering is to group similar objects together, it was understood that the results for nonsingleton clusters only were more valid. The objective of varying the parameters *p* and *m* was, in part, aimed not only to reduce the number of singleton clusters but also to see the effects on the performance measure.

Table 6. Variation in Performance of Coefficients with Jarvis-Patrick Parameters a. BCI Fingerprints, b. Daylight Fingerprints, c. Unity Fingerprints, and d. ECFP4 Fingerprints

cluster parameters	Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
a. BCI Fingerprints													
All Nonsingleton Clusters													
14, 6	11	0	11	11	11	11	0	11	0	10	9	11	11
14, 8	11	0	11	11	11	11	0	11	0	11	0	11	11
16, 6	0	3	0	0	0	1	0	0	7	0	9	0	0
16, 8	11	0	10	11	11	11	0	11	0	11	1	11	11
20, 6	0	2	0	0	0	0	0	0	10	0	0	0	0
20, 8	0	6	0	0	0	0	1	0	10	0	1	0	0
All Clusters, Including Singletons													
14, 6	0	5	0	0	0	0	4	0	6	0	0	0	0
14, 8	0	4	0	0	0	0	6	0	9	0	1	0	0
16, 6	0	4	0	0	0	0	2	0	6	0	0	0	0
16, 8	0	3	0	0	0	0	4	0	9	0	0	0	0
20, 6	0	7	0	0	0	0	1	0	6	0	0	0	0
20, 8	0	4	0	0	0	0	1	0	8	0	0	0	0
b. Daylight Fingerprints													
All Nonsingleton Clusters													
14, 6	11	0	3	10	11	11	0	11	0	11	4	11	10
14, 8	11	0	11	11	11	11	0	11	0	11	0	11	11
16, 6	0	11	1	0	0	0	0	0	11	0	0	0	0
16, 8	11	0	8	11	11	11	0	11	0	11	1	11	11
20, 6	0	7	0	0	0	0	0	0	11	0	0	0	0
20, 8	0	10	0	0	0	0	0	0	11	0	0	0	0
All Clusters, Including Singletons													
14, 6	0	2	0	0	0	0	5	0	6	0	0	0	0
14, 8	0	1	0	0	0	0	9	0	10	0	0	0	0
16, 6	0	2	0	0	0	0	3	0	7	0	0	0	0
16, 8	0	0	0	0	0	0	8	0	8	0	0	0	0
20, 6	0	1	0	0	0	0	9	0	5	0	0	0	0
20, 8	0	1	0	0	0	0	2	0	8	0	0	0	0
c. Unity Fingerprints													
All Nonsingleton Clusters													
14, 6	11	0	10	10	11	11	0	10	0	11	1	11	11
14, 8	11	0	11	11	11	11	0	11	0	11	0	11	11
16, 6	0	0	0	0	0	0	0	0	0	0	11	0	0
16, 8	11	0	11	11	11	11	0	11	0	11	0	11	11
20, 6	0	0	0	0	0	0	0	0	11	0	0	0	0
20, 8	0	3	0	0	0	0	0	0	11	0	7	0	0
All Clusters, Including Singletons													
14, 6	0	6	0	0	0	0	5	0	6	0	0	0	0
14, 8	0	4	0	0	0	0	6	0	11	0	0	0	0
16, 6	0	6	0	0	0	0	3	0	3	0	0	0	0
16, 8	0	5	0	0	0	0	5	0	7	0	0	0	0
20, 6	0	4	0	0	0	0	1	0	6	0	0	0	0
20, 8	0	4	0	0	0	0	1	0	6	0	0	0	0
d. ECFP4 Fingerprints													
All Nonsingleton Clusters													
14, 6	11	0	10	11	11	11	10	11	0	11	11	11	11
14, 8	11	0	11	11	11	11	11	11	0	11	11	11	11
16, 6	3	5	0	2	6	8	0	8	7	5	1	5	2
16, 8	11	0	11	11	11	11	11	11	0	11	11	11	11
20, 6	0	11	0	0	0	0	0	0	0	0	0	0	0
20, 8	7	5	0	4	10	5	1	10	8	9	7	8	2
All Clusters, Including Singletons													
14, 6	0	7	3	0	0	0	5	0	8	0	0	0	0
14, 8	3	6	6	4	3	4	7	3	9	3	4	3	4
16, 6	0	6	1	0	0	1	0	0	3	0	1	0	0
16, 8	2	7	4	1	2	2	6	2	10	2	1	2	2
20, 6	0	11	0	0	0	0	0	0	0	0	0	0	0
20, 8	0	4	1	0	0	0	1	0	7	0	0	0	0

DISSIMILARITY-BASED COMPOUND SELECTION

The aim of dissimilarity-based compound selection, as we have mentioned, is to select a maximally diverse set of

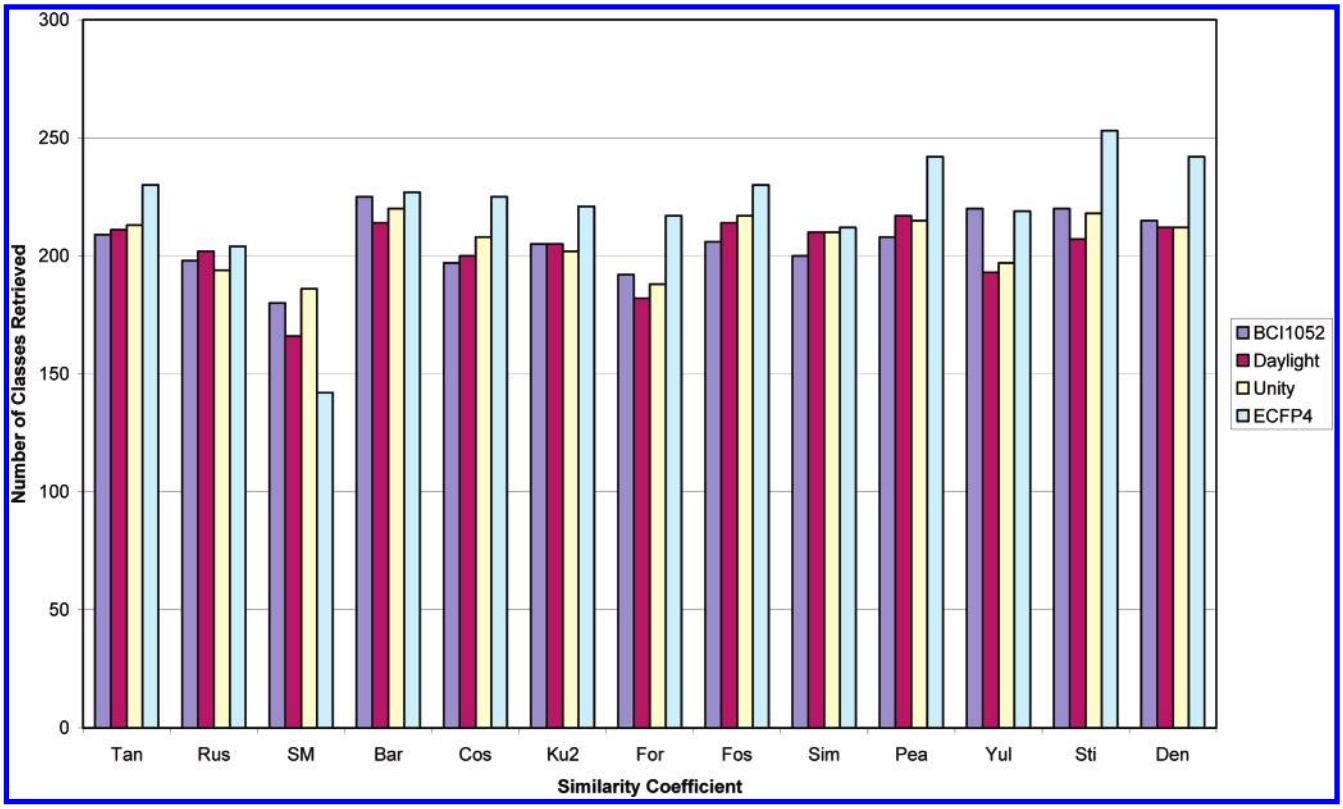

Figure 3. Classes identified using dissimilarity-based compound selection.

compounds in order that the widest range of structural variation, and therefore biological activity, can be represented. We have used a standard algorithm, outlined in Figure 1, to select a subset of compounds from the 20K data set. The first compound selected is that which is most dissimilar to all other compounds. The data set represents a total of 591 active classes, with some compounds being active in more than one class. The best-case scenario for our compound selection routine would involve maximum class coverage represented by as few selections as possible. With this aim in mind, we decided to select 591 compounds and took our measure of performance as the number of different classes represented by this subset. It is possible that the 591 classes could be represented by a smaller number of compounds, but that this is almost certainly unattainable via automated means. Again, each of the 13 coefficients of Table 1 was used as the similarity metric, and the routine was repeated for each fingerprint representation.

HIERARCHICAL CLUSTERING RESULTS

The 20K data set was characterized using each of the four fingerprints described. Each of these characterizations was then clustered 13 times using the group-average agglomerative hierarchical clustering method with one of the coefficients of Table 1 applied as the similarity metric in each case. Samples of the clusters were taken at every hundredth iteration of the clustering routine so that evaluation of the relative performance for each of the 11 active classes could be carried out. The performance measure described above was applied to these samples, and the results were plotted for each active class. Figure 2 shows the results for the BCI1052 characterizations of the angiotensin II AT1 antagonist class with the last 2000 cluster levels enlarged. Focusing on the last 2000 cluster levels, there appear to be four poor-performing coefficients, the Simpson, Simple Match, Forbes, and, worst of all, the Russell/Rao. A similar pattern resulted with most of the other classes, with the exception of substance P antagonists, for which the Simple Match performed comparably to other coefficients, and renin inhibitors, for which the Simple Match was the best performing coefficient.

Table 3a−c illustrates the performance measure values for all classes and coefficients at levels of 2000 clusters, 1000 clusters, and 500 clusters (BCI1052). The tables show the relative performance of the coefficients for each class, with the best performer shown in italic boldface and those performing within 10% of the best performer in boldface. Clearly, the poor performers are the Russell/Rao, Forbes, Simpson, and, with the exception of the renin inhibitors, the Simple Match. This is supported by the final row of each table which indicates the number of boldface entries. Taking the sum of values across all three tables, several coefficients are clearly superior in a clustering technique of this type, these being the Pearson, Stiles, Baroni-Urbani/Buser, and Yule. Noticeably, three of these are correlation coefficients. Similar levels of performance are observed when using the other three fingerprint types, as shown in Table 4 which illustrates the number of boldface (i.e., within 10% of the best performer) entries for each characterization at levels of 2000 clusters, 1000 clusters, and 500 clusters.

NONHIERARCHICAL CLUSTERING RESULTS

The data set of 20K MDDR compounds was clustered using the Jarvis-Patrick clustering algorithm using various strategies. All four fingerprint characterizations were applied to each clustering strategy. Clustering was carried out using

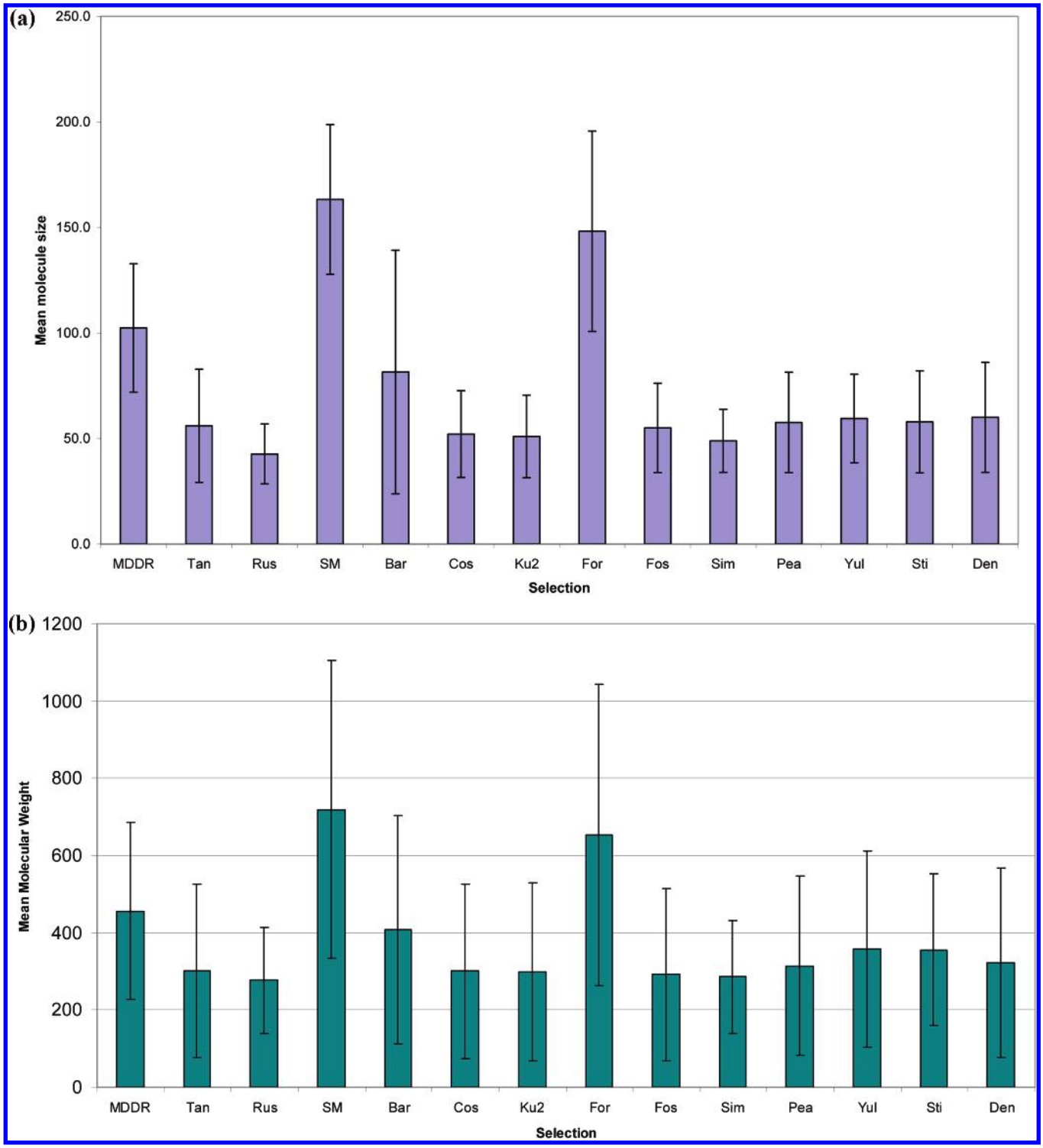

Figure 4. a. Mean size of molecules selected in terms of bit density, with standard deviations (BCI1052) and b. mean molecule weights for compounds selected, with standard deviations.

each of the 13 coefficients of Table 1 as the similarity metric. The clustering parameters were varied, using nearest neighbor list lengths (m) of 14, 16, and 20 and using values of 6 and 8 for the minimum common nearest neighbor requirement (p). Table 5a,b shows the results, given by our performance measure, for parameters $m=14$, $p=8$ and $m=20$, $p=8$, and using the BCI1052 fingerprints, in which all singleton clusters have been omitted. This table, like that of Table 3, indicates the best-performing coefficient for each class in italic boldface and those that perform within 10% of the best performer in boldface. The final row indicates the number of times a coefficient appears as a boldface entry. Similar results are obtained with the other three fingerprint types. Table 6a−d summarizes the results for all parameter values tested for the four fingerprint types, indicating the number entries which are within 10% of the best performer (cf. boldface entries of Table 5).

When applying the standard parameters of $m=14$, $p=8$, the nonsingleton clusters indicate clearly that the Russell/Rao, Forbes, Simpson, and Yule coefficients are not suitable when applied to this clustering technique. The other nine coefficients exhibit comparable performance. Increasing the

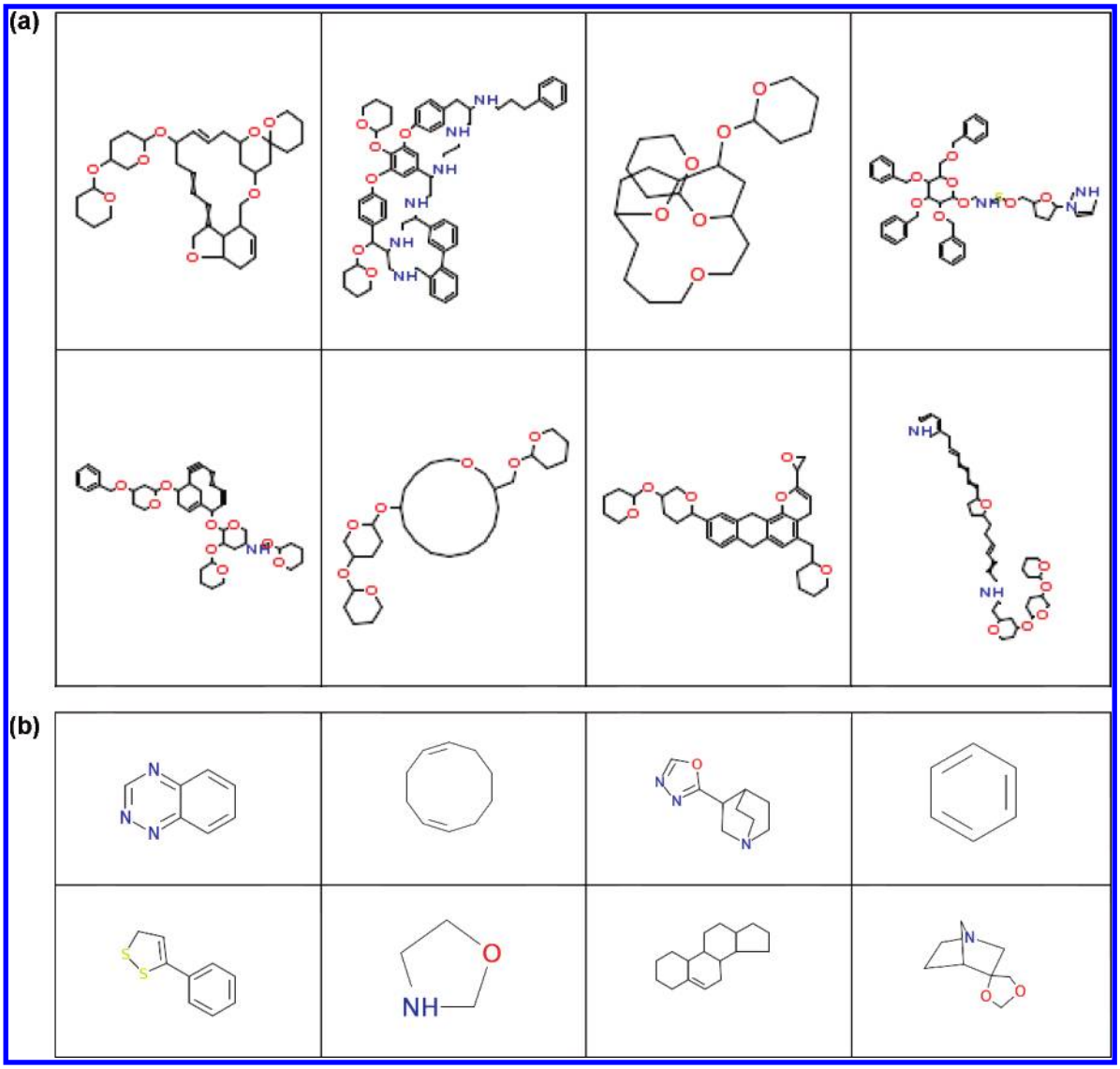

Figure 5. a. Example scaffolds from the Simple Match subset and b. example scaffolds from the Russel/Rao subset.

nearest neighbor list length has the effect of increasing the similarity threshold for those clusters containing the bulk of the data set, thus isolating those clusters which contain the outliers. The results are therefore more akin to those which include singletons, as the performance measure is reduced for the larger and less tightly defined clusters, but remains high for the smaller and singleton clusters. Hence the performance tends to improve for those coefficients which are known to identify the outliers, such as Russell/Rao, which is biased toward larger compounds, and Forbes, which is biased toward smaller compounds.

DISSIMILARITY-BASED COMPOUND SELECTION RESULTS

Figure 3 illustrates the number of classes identified when the 13 coefficients are applied to the compound selection strategy, described in Figure 1, in which 591 compounds are selected. The results are less clear-cut, although they do identify poorer performing coefficients, such as the Forbes, Russell/Rao, and, more clearly, the Simple Match. With the exception of these, the 591 active classes are well represented. The size range of compounds selected is, however, limited in all cases. Figure 4a,b illustrates the sizes of compounds being selected for each coefficient in terms of bit density and molecular weight as well as those of the full 20K data set, when using the BCI1052 representation. As has previously been noted, the Simple Match, effectively the complement of the Squared Euclidean Distance, has a tendency to select larger (in terms of bit density) compounds, while the Tanimoto selects smaller compounds. Indeed, most of the coefficients tested appear to select compounds which are smaller than the database average. Only the Simple Match and Forbes select the larger compounds. This is an expected reversal of the trend observed in similarity searching[5] in which the Russell/Rao and Tanimoto are seen to retrieve larger compounds and the Forbes and Simple Match retrieve smaller ones.

Figure 4 also illustrates the standard deviation in the size of the compounds selected. The main purpose of compound selection is to identify a wide structural variety of compounds but with a similar distribution to the full database. The closest mean value to the full database would appear to be that of the Baroni-Urbani/Buser,[23] and this also has a wide distribution as shown by the standard deviation. Figure 3 also

Table 7. Number of Unique Murcko Scaffolds Identified for the Compounds Selected (BCI1052)

Tan	Rus	SM	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
327	276	512	371	309	314	484	338	305	327	341	286	334

indicates a good representation of the 591 classes for this coefficient. These features would appear to indicate that the Baroni-Urbani/Buser is most suitable for compound selection, although Z-test statistics indicate that all subset means are well outside the MDDR distribution.

Clearly, the $\sqrt{ab}$ term in the Baroni-Urbani/Buser coefficient has a normalizing effect on coefficients biased toward common presence as well as those biased toward common absence. A similar effect is expected in the Modified Tanimoto coefficient,[24] which includes a contribution from common unset bits in the bitstring.

Within the various size ranges being selected, the variety of compounds in each subset is, as would be expected, considerable. Table 7 indicates the number of unique Murcko scaffolds (calculated using Scitegic's Pipeline Pilot) identified in each of the 591 compound subsets. Notably, the Simple Match has the largest variety of scaffolds, with 512 represented by the subset. This is not surprising given the size of these compounds (see Figure 4). With the compounds being so large, there is more scope for structural variation as can be seen by the selection of Figure 5a. The subset selected by Russel/Rao, in contrast, contains scaffolds which are mostly single rings, linked single rings, or single fused ring systems (a selection is illustrated in Figure 5b).

DISCUSSION

The clustering results above appear to suggest a preference for those coefficients which are not size-biased. Coefficients such as the Simple Match, Russell/Rao, or Forbes, when applied to similarity search strategies, show a bias toward size distributions which differ from those of the database. For clustering methodologies, it would appear that the imposition of size-related grouping onto the clustering procedure reduces the chemical meaning of purely structure-based classification. Since the procedure is applied to the whole data set and is not aimed specifically at a single class or target as we find in similarity searching, the classification benefits from the removal of this bias. Even for classes which are likely to have a size distribution which differs from the full database, the influence of size-biased coefficients does not appear to have any beneficial effect on the classification.

These effects are illustrated in compound selection, where the size-bias is clearly detrimental to the selection process, reducing the range of classes being selected as a result of the reduction in the size range from which compounds can be selected. Two aspects of this study are apparent. First, that all of the coefficients commonly applied to compound selection routines, notably the Tanimoto and Simple Match (cf. Euclidean Distance), suffer from the effects of size bias. Second, that only one coefficient, the Baroni-Urbani/Buser, selects compounds which exhibit a similar size distribution to the database, although not a significant similarity distribution in statistical terms. If well-known coefficients such as the Tanimoto and Euclidean Distance are to be applied, performance is likely to be improved by combining the selections of complementary coefficients, either by merging the selections or via some intelligent data fusion technique or by employing a combined similarity coefficient such as that of Fligner et al.,[24] thus normalizing the bias effect.

It would appear from the results of Tables 4 and 6 and Figure 3 that the choice of fingerprint representation has little significant effect on the results. Although the BCI fingerprints have been exemplified in the text, figures, and tables, similar results were obtained for the Daylight, Unity, and Scitegic characterizations.

ACKNOWLEDGMENT

We acknowledge support from the British Council and the Polish Ministry of Science and Education for funding the visit of Maciej Haranczyk to the University of Sheffield through the Young Scientists Programme (WAR/342/86). This work was also supported by the Polish State Committee for Scientific Research (KBN) Grants DS/8000-4-0026-8 (M.H.) and N204 127 31/2963 (M.H.). Maciej Haranczyk is a holder of the award from the Foundation for Polish Science (FNP). We also acknowledge Daylight Chemical Information Systems Inc., Digital Chemistry, Tripos Inc. and Scitegic for the provision of fingerprinting software and MDL (now Symyx Technologies Inc.) for the MDDR database.

REFERENCES AND NOTES

(1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity;* John Wiley and Sons: New York, 1990.
(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(3) Willett, P. Similarity Methods in Chemoinformatics. *Annu. Rev. Inf. Sci. Technol.* In press.
(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 Screening* **2002**, *5*, 155−166.
(5) 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;* Lavine, B., Ed.; American Chemical Society: Washington, DC, 2005; Vol. 894, pp 77−95.
(6) 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.
(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) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.
(9) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.
(10) Willett, P. *Similarity and clustering in chemical information systems*; Research Studies Press: Letchworth, 1987.
(11) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Statist. Assoc.* **1963**, *58*, 236−244.
(12) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(13) Lajiness, M. S. Molecular similarity-based methods for selecting compounds for screening. In *Computational chemical graph theory;* Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 299−316.
(14) Bawden, D. Molecular dissimilarity in chemical information systems. In *Chemical structures 2;* Warr, W. A., Ed.; Springer-Verlag: Heidelberg, 1993; pp 383−388.
(15) The MDL Drug Data Report, available from Symyx Technologies. http://www.mdli.com/products/knowledge/drug_data_report (accessed Dec 17, 2007).
(16) Digital Chemistry. http://www.digitalchemistry.co.uk (accessed Dec 17, 2007).

(17) Daylight Chemical Information Systems, Inc. http://www.daylight.com (accessed Dec 17, 2007).
(18) Tripos, Inc. http://www.tripos.com (accessed Dec 17, 2007).
(19) Scitegic. http://www.scitegic.com (accessed Dec 17, 2007).
(20) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry;* Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley: Hoboken, NJ, 2002; Vol. 18, pp 1−40.
(21) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E., et al. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(22) Mojena, R. Hierarchical grouping methods and stopping rules: an evaluation. *Comput. J.* **1997**, *20*, 359−363.
(23) Baroni-Urbani, C.; Buser, M. W. Similarity of binary data. *Syst. Zool.* **1976**, *25*, 251−259.
(24) 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.

CI700413A