Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion

Naomie Salim, John Holliday,* and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Received September 2, 2002

Many different types of similarity coefficients have been described in the literature. Since different coefficients take into account different characteristics when assessing the degree of similarity between molecules, it is reasonable to combine them to further optimize the measures of similarity between molecules. This paper describes experiments in which data fusion is used to combine several binary similarity coefficients to get an overall estimate of similarity for searching databases of bioactive molecules. The results show that search performances can be improved by combining coefficients with little extra computational cost. However, there is no single combination which gives a consistently high performance for all search types.

INTRODUCTION

Calculation of the similarity or dissimilarity between two molecules is a standard practice in the chemoinformatics field. Such calculations are used in similarity searching,[1] property prediction,[2] synthesis design,[3] virtual screening,[4] and molecular diversity analysis[5] inter alia. For example, techniques for selecting a subset of maximally diverse compounds from a large database[6] involve the calculation of similarity between each selected molecule and every other molecule in the database.

Two principal factors which affect the performance of these calculations are the representation used to characterize the molecules and the similarity coefficient used to compare them. The representation is usually either a fingerprint, essentially a set of binary elements describing the presence or absence of predefined substructural fragments, or a set of calculated physicochemical properties.[7,8] The coefficient can be drawn from a varied selection which has been described in the literature, most of which can be grouped into three broad classes: association coefficients, correlation coefficients, and distance coefficients.

Association coefficients are commonly used with binary representations and are often normalized to lie within the range of zero (no similar features in common) and unity (identical representations). However, they can be used with nonbinary representations, in which case the range may be different. Correlation coefficients measure the degree of correlation between sets of values characterizing a pair of objects. Distance coefficients quantify the degree of dissimilarity between two objects and, when normalized and using binary data, range between zero (identity) and unity (no similar features in common).

In a previous study[9] similarity coefficients were examined in order to group together those with comparable performance when applied to searches of binary representations. The result of this study was a subset of coefficients which exemplified the full range of types of coefficient. Example coefficients from each of the groups were then used to determine whether the performance of single coefficients could be improved by combining coefficients using data fusion. To evaluate their performance, sets of active targets were used in a series of similarity searches against two test databases, the 1999 version of the NCI AIDS database[10] and the IDAlert database from Current Drugs.[11] The binary representations used were UNITY 2D fragment bit-strings,[12] hashed fingerprints which encode the 2D structural features present in the molecule.

This paper reports on further studies which have extended the evaluation to include MDL's MDDR database[13] and have also used alternative representations on all test databases: these being Barnard Chemical Information (BCI) standard bit-strings[14] and Daylight fingerprints.[15] In addition, the factors determining which coefficients one might choose in combination are investigated. Results show that there is never one true "best performer" and that the relative performance of the coefficient combinations is related to the characteristics of the data itself.

CLUSTERING COEFFICIENTS

The set of 22 similarity coefficients, shown in Figure 1, was selected from an extensive review by Ellis et al.[16] The formulas shown in Figure 1 indicate the similarity (or dissimilarity in the case of coefficient 22, the Mean Manhattan) between two bit-string representations, X and Y, of length n, in which a is the number of bits set in both X and Y, b is the number of bits set exclusively in X, c is the number of bits set exclusively in Y, and d is the number of bits set in neither X or Y. In the case of the distance coefficient (Mean Manhattan), since the coefficient values range from zero (identity) to one (no bits in common), the complement of the coefficient value defines its similarity value.

For each test database, several query structures were selected to be used in similarity searches. These similarity searches were carried out against the respective test database using all 22 coefficients separately. The database molecules were ranked in decreasing order of the calculated similarity coefficient. The rankings of two coefficient searches for the same query can be compared by counting the number of

* Corresponding author e-mail: j.d.holliday@sheffield.ac.uk.

10.1021/ci025596j CCC: $25.00
Published on Web 12/20/2002

1. Jaccard/Tanimoto $\frac{a}{a+b+c}$
2. Dice $\frac{2a}{2a+b+c}$
3. Russell/Rao $\frac{a}{n}$
4. Sokal/Sneath(1) $\frac{a}{a+2b+2c}$
5. Kulczynski(1) $\frac{a}{b+c}$
6. Simple Matching $\frac{a+d}{n}$
7. Hamann $\frac{a+d-b-c}{n}$
8. Sokal/Sneath(2) $\frac{2a+2d}{a+d+n}$
9. Rogers/Tanimoto $\frac{a+d}{b+c+n}$
10. Sokal/Sneath(3) $\frac{a+d}{b+c}$
11. Baroni-Urbani/Buser $\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
12. Ochiai/Cosine $\frac{a}{\sqrt{(a+b)(a+c)}}$
13. Kulczynski(2) $\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$
14. Forbes $\frac{n\times a}{(a+b)(a+c)}$
15. Fossum $\frac{n\left(a-\frac{1}{2}\right)^2}{(a+b)(a+c)}$
16. Simpson $\frac{a}{\min(a+b,a+c)}$
17. Pearson $\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
18. Yule $\frac{ad-bc}{ad+bc}$
19. McConnaughey $\frac{a^2-bc}{(a+b)(a+c)}$
20. Stiles $\log_{10}\frac{n\left(\lvert ad-bc\rvert-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$
21. Dennis $\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$
22. Mean Manhattan $\frac{b+c}{n}$

Figure 1. Association coefficients (1−16), correlation coefficients (17−21), and distance coefficient (22) used in the experiments.

compounds in common in the top t ranked structures, where t was initially chosen as 400. The dissimilarity between two rankings for coefficients i and j is then defined as

$$D_{ij} = 1 - \frac{c_{ij}}{t}$$

where c_{ij} is the number of common structures in the top t ranking positions. For each query, a 22 × 22 dissimilarity matrix was generated using all of the coefficients. The dissimilarity matrix was then clustered using three hierarchical clustering methods, single linkage, complete linkage, and group-average, which resulted in three respective dendrograms.

In order that groupings of related coefficients are formally identified, Mojena's stopping rule[17] was then applied to the dendrograms. Mojena's stopping rule uses the distribution of clusters to identify when "a significant change from one stage to the next implies a partition which should not be undertaken"; effectively a formal rule for determining a natural cutoff where the clusters are at an optimum. When applied to the dendrogram of Figure 2, Mojena's stopping rule identifies three groups as follows (the numbers in the groups correspond to the coefficient numbers in Figure 1):

Group A: {1 2 4 5 11 12 15 17 20 21}

Group B: {3 13 16 18 19}

Group C: {6 7 8 9 10 14 22}

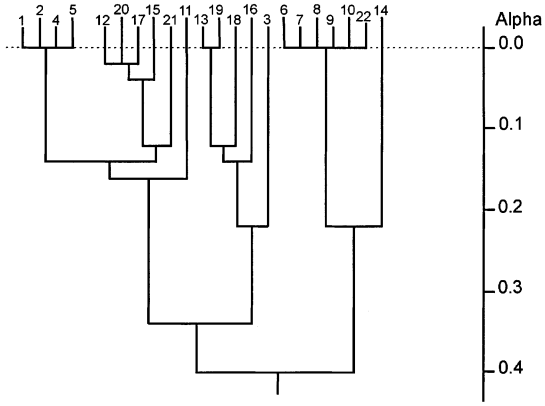

Figure 2. Dendrogram for single linkage clustering of first target structure dissimilarity matrix; $\alpha_{j+1} = 0.274$ implies stopping at level 19 by Mojena's rule ($k = 1.25$).

The previous study[9] reported the results of 60 searches against the AIDS database and 60 searches against the IDAlert database, using values of 50, 100, 200, and 400 for the parameter t, in which the molecules were characterized by UNITY 2D bit-strings. When all of the dendrogram groupings were combined, a total of 11 groups of related coefficients was identified. The coefficients within each of these 11 groups consistently clustered into the same groups based on the Mojena stopping rule.

Further detailed evaluation has been carried out using equivalent searches where the AIDS and IDAlert databases

were characterized by BCI bit-strings and Daylight fingerprints and using all three binary representations for similarity searches on the MDDR database. This extensive study identified a total of 13 groups, where previously the Pearson (17) and Stiles (20) coefficients had been grouped together, as had the Ochiai/Cosine (12) and Fossum (15) coefficients. The Pearson and Stiles were grouped together in all but a few of the dendrograms, as were the Ochiai/Cosine and Fossum, so the results were very similar to the previous study. We have maintained their separate groupings here for completeness, giving the 13 groups as follows:

Group A: {3}
Group B: {11}
Group C: {14}
Group D: {16}
Group E: {18}
Group F: {21}
Group G: {1 2 4 5}
Group H: {6 7 8 9 10 22}
Group I: {12}
Group J: {15}
Group K: {13 19}
Group L: {17}
Group M: {20}

INITIAL DATA FUSION STUDIES

Similarity measures in the chemical information field have, in the main, been limited to a few single measures such as the Tanimoto, Cosine, and Euclidean Distance. Indeed, many of those shown in Figure 1 have rarely been used in connection with chemical structure similarity calculations. Several have been used in comparative studies, seeking to determine which is the best single measure for performing a particular task, a similarity search for instance.[18] The variety of performance of the similarity measures available, as shown above, has led to an interest in data fusion methods for combining more than one similarity measure with a goal to improving the performance of the single measure. A study by Ginn et al.[19] suggests that combining the rankings of searches using more than one coefficient would give improved performance, and this was further tested by Holliday et al.[9] This section reports on further detailed experiments which aim to investigate the factors which determine the choice of coefficients for data fusion.

We have shown that there is considerable variety in the types of coefficient available for measuring similarity and have categorized these into the 13 groups shown above. We selected one coefficient from each of the groups in order to represent the full range of types of similarity measure available. These were as follows: (3) **Rus**sell/Rao, (11) **Bar**oni-Urbani/Buser, (14) **For**bes, (16) **Sim**pson, (18) **Yul**e, (21) **Den**nis, (1) Jaccard/**Ta**nimoto, (6) **S**imple **M**atch, (12) Ochiai/**Cos**ine, (15) **Fos**sum, (13) **Ku**lczynski **2**, (17) **Pea**rson, and (20) **Sti**les. For the remainder of this paper, we will use a code (shown in bold) to denote each coefficient.

We selected active targets from seven activity types (5HT4 agonist, adrenergic β, dopamine agonist, ACE inhibitors, HIV-1 protease inhibitors, benzodiazepine agonists, and lactamase (beta) inhibitors) to carry out several searches on the MDDR database, characterized by UNITY 2D bit-strings, BCI standard bit-strings, and Daylight fingerprints in a comparative study to find the best combination of similarity measures.

Fusion was based on a summing procedure on the rankings produced by the searches. This was found to be the most effective method in the study of Ginn et al. For each database compound, its rank position was summed across all of the rankings for coefficients involved in the fusion. (E.g. if a database compound appeared 20th, 15th, and 7th in the rankings for searches using three coefficients, it would have a sum of 42). This was repeated for all database compounds, and a ranking was produced by ordering the database compounds based on their sums. The combined ranking was examined to identify compounds which exhibited the same activity as the target in the top 5% rank positions. This number was used as a measure of the effectiveness of the respective combination of similarity measures. The number of coefficients fused ranged from 1, the single measures, to all 13. All of the ${}_nC_{13}$ possible combinations for $n = 1$ to 13 were investigated. For comparative purposes, for each value of n, the combinations were sorted into decreasing order of actives retrieved, and each combination was then given an ordinal number based on this ordering (ties were given an average ordinal number). The overall performance of each combination was taken as the sum of ordinal values obtained for that combination across all searches.

Figure 3 shows the performance of the combinations, in terms of average number of actives retrieved in the top 5% of rankings over all targets for five of the activity types (using UNITY 2D bit-strings). All but the lactamase (beta) inhibitors show an increase in performance over single measures when using fused combinations. This performance increase seems to peak when about two to four coefficients are combined, but, in many cases, still remains high for nearly all coefficients in combination. Similar results are obtained using Daylight fingerprint and BCI bit-string characterizations. The average performance of fused combinations rarely exceeds the performance of the best single coefficient, which would indicate that it is important to choose the right coefficients in a fused combination.

To decide which are the preferred coefficients for a combination, we can look at their relative merits in the combinations in which they are involved. For example, each coefficient will occur in 66 of the ${}_3C_{13}$ 3-coefficient combinations. The sum of the ordinal values for all combinations containing a coefficient will reflect the overall performance of that coefficient for that particular search and therefore the desirability for its use in a fusion combination. Table 1 shows the comparative performance based the sum of ordinal values for 2-, 3-, and 4-coefficient combinations involving searches for 5HT4 agonist targets. Very similar results were obtained in all searches for HIV-1 protease inhibitors and benzodiazepine agonists and a few searches involving lactamase (beta) inhibitors. These results would suggest choosing Tanimoto and Russell/Rao for a 2-coefficient combination, Russell/Rao, Tanimoto, and Fossum for a 3-coefficient combination, and Russell/Rao, Tanimoto, Fossum, and Cosine for a 4-coefficient combination.

However, looking at Table 2, which illustrates the relative merits when applied to searches for dopamine agonists using

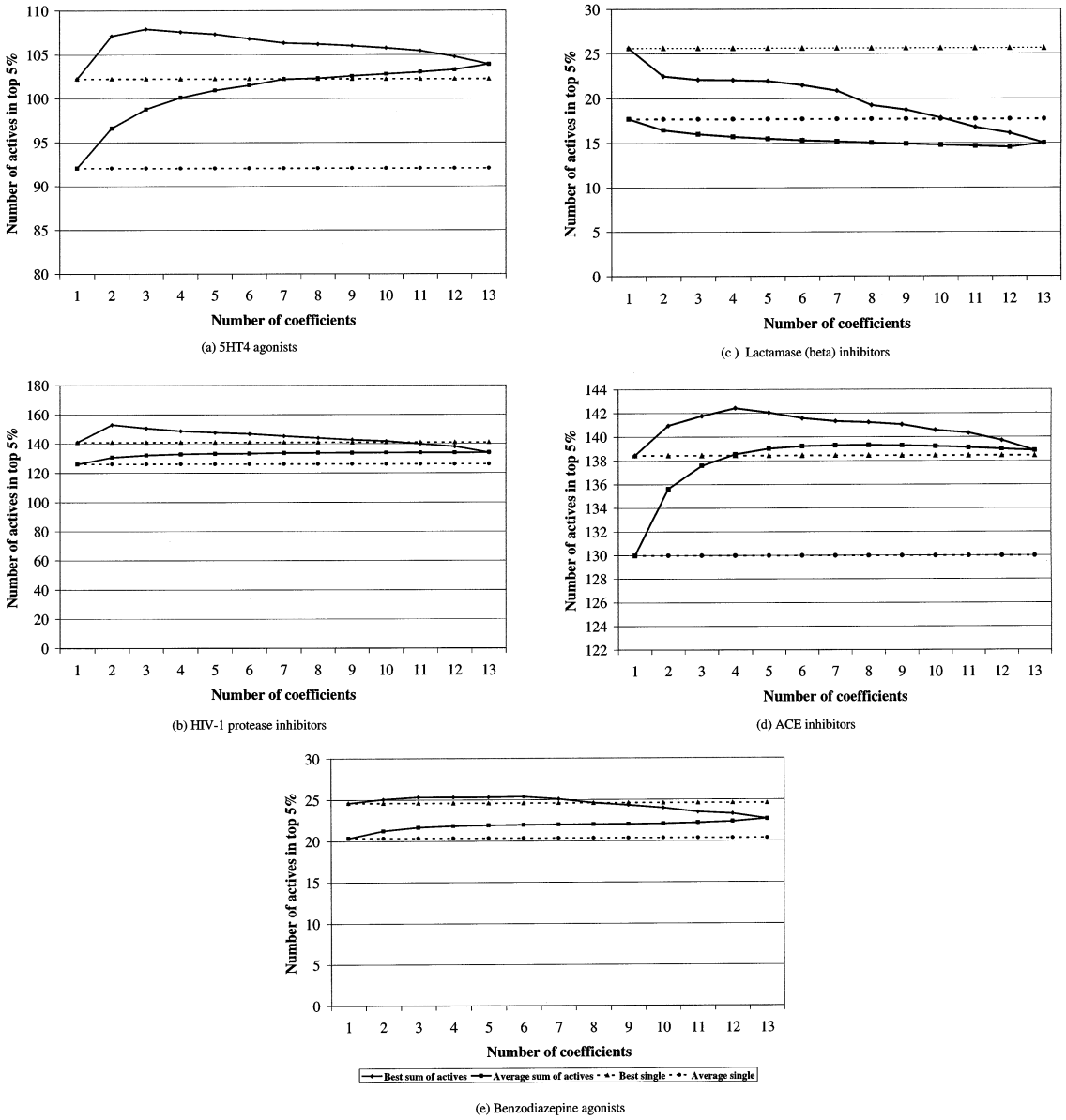

Figure 3. Performance of fusions of coefficients compared to performance of single coefficient, using UNITY 2D bit-strings.

Table 1. Sum of Rank Positions for Combinations Involving Each Coefficient (for 5HT4 Agonists, UNITY 2D Bit-Strings)

(a) performance in 2-fusions			(b) performance in 3-fusions			(c) performance in 4-fusions		
coefficient	sum of rank	position	coefficient	sum of rank	position	coefficient	sum of rank	position
Tan	8605.5	1	Rus	172568	1	Rus	1632559	1
Rus	9016.5	2	Tan	240528.5	2	Tan	2637350.5	2
Fos	9627	3	Fos	266571.5	3	Fos	2854691	3
Cos	9661.5	4	Cos	267638	4	Cos	2860891.5	4
Bar	9992	5	Bar	274872	5	Bar	2918762.5	5
Pea	11106	6	Ku2	295608	6	Ku2	3038777.5	6
Ku2	11553	7	Pea	295684.5	7	Pea	3053630	7
Sti	11880	8	Den	304531.5	8	Den	3109162.5	8
Den	12305.5	9	Sti	314208.5	9	Yul	3220520	9
Yul	13699	10	Yul	322462.5	10	SM	3226044	10
SM	14659.5	11	SM	331960.5	11	Sti	3235104.5	11
Sim	17137	12	Sim	356293.5	12	Sim	3385133	12
For	18088	13	For	387270.5	13	For	3643330	13

BCI bit-strings, the choice would be very different. The Russell/Rao would appear to be a very poor choice, and the Tanimoto, Fossum, and Cosine would not be included in a preferred combination. Very similar results were obtained for searches involving ACE inhibitors.

DISCUSSION OF INITIAL STUDIES

Searches involving other targets and using alternative characterizations yielded a considerable range of preferred coefficients, indicating that there is no one combination

Table 2. Sum of Rank Positions for Combinations Involving Each Coefficient (for Dopamine Agonists, BCI 1052 Bit-Strings)

(a) performance in 2-fusions			(b) performance in 3-fusions			(c) performance in 4-fusions		
coefficient	sum of rank	position	coefficient	sum of rank	position	coefficient	sum of rank	position
For	7200.5	1	For	214437	1	For	2575991.5	1
SM	8044.5	2	SM	215191.5	2	SM	2619884	2
Bar	10112.5	3	Bar	262500.5	3	Bar	2849469.5	3
Yul	11745.5	4	Den	293864	4	Den.	2963804.5	4
Den	11997.5	5	Yul	294020.5	5	Yul	2984023.5	5
Tan	12350.5	6	Tan	296749	6	Pea	3006111	6
Pea	12630	7	Pea	302956	7	Tan	3010483.5	7
Cos	12847.5	8	Cos	303659	8	Cos	3047224	8
Fos	13073.5	9	Fos	304222	9	Fos	3050187.5	9
Ku2	13730.5	10	Ku2	312172	10	Ku2	3054727.5	10
Sti	14357.5	11	Sti	344569.5	11	Sim	3166975.5	11
Sim	15950.5	12	Sim	353051.5	12	Sti	3302855.5	12
Rus	19905	13	Rus	449574	13	Rus	3790378.5	13

Table 3. Best Coefficients and Best Combinations for 2- to 4-Coefficient Fusions, with Actives Sorted by Their Average Number of Bits Set

bit string	active	av bit set	best single	best in 2-fusion	best in 3-fusion	best in 4-fusion	best 2-fusion	best 3-fusion	best 4-fusion	best fusion better than best single
UNITY 2D	ACE inhibitor	211	Sti	Den	Den	For	Bar,Yul	Bar,Yul, Den	Rus,SM, For,Sim	yes
UNITY 2D	5HT4 agonist	214	Tan	Tan	Rus	Rus	Rus,For	Rus,For, Pea	Rus,SM, Yul,Den	yes
UNITY 2D	HIV-I protease inhibitor	251	Rus	Rus	Rus	Rus	Rus,Bar	Rus,Ku2, Fos	Tan,Rus, Bar,Pea	yes
UNITY 2D	benzodiazepine agonist	261	Tan	Rus	Rus	Rus	Rus,For	Rus,Pea, Den	Rus,Bar, Ku2,Yul	yes
UNITY 2D	lactamase(beta) inhibitor	281	Rus	Rus	Rus	Rus	Rus,Fos	Rus,Cos/Fos,Ku2	Rus,Cos, Ku2,Fos	no
BCI 1052	dopamine agonist	82	For/SM	For	For	For	Bar,For	SM,For, Den	Tan,SM, Bar,For	yes
BCI 1052	ACE inhibitor	105	Den	For	For	For	For,Fos	Tan,Ku2, For	Bar,Cos, For,Fos	yes
BCI 1052	benzodiazepine agonist	105	Tan	Rus	Rus	Rus	Rus,For	Rus,Ku2, For	Tan,Rus,Cos/Fos, For	yes
BCI 1052	lactamase(beta) inhibitor	111	Pea/Yul	Bar	Tan	Ku2	Tan,Bar	SM,Bar, Fos	Tan,SM,Cos/Ku2,Fos	yes
BCI 1052	HIV-protease inhibitor	120	Tan	Rus	Rus	Rus	Rus,Ku2	Tan,Rus,Fos	Tan,Rus,Bar,Pea	no
Daylight	adrenergic β agonist	233	Pea	Sti	Fos	Sim	Bar,Sim	Tan,Sim,Sti	Rus,SM,Sim,Yul	yes
Daylight	ACE inhibitor	269	Ku2	Ku2	Sim	Sim	Fos,Sim	Ku2,Sim,Pea	BarKu2,Sim,Yul	yes
Daylight	HIV-I protease inhibitor	312	Tan	Rus	Rus	Rus	Rus,Cos/Fos	Tan,Rus,Fos	Rus,Ku2,Yul,Cos/Fos	yes
Daylight	lactamase(beta) inhibitor	355	Rus	Ku2	Ku2	Ku2	CosFos	Cos,Ku2,Fos	TanBar,Cos,Fos	no
Daylight	benzodiazepine agonist	364	Fos	Rus	Rus	Rus	Rus,Den	Rus,Fos,Pea	Rus,Bar,Cos/Fos,Sim	yes

which would be a standard choice and that there may be factors which influence the choice of coefficient. The results do, however, indicate that, in most cases, a combination will outperform the best single coefficient.

The possibility that the choice is size-based is reflected in Table 3 which shows the best performing single and combined coefficients for 2- to 4-coefficient fusions. These results are ordered by the average number of bits set (or average bit density) for each activity type, generally relating to the average size of the molecules in terms of the number of constituent atoms.

There is a noticeable preference for certain coefficients to perform well in certain size ranges. Russell/Rao and Fossum, for example, appear in many of the best combinations involving large targets such as the HIV-1 protease inhibitors and benzodiazepine agonists, whereas other coefficients such as Forbes and Baroni-Urbani/Buser appear in combinations which perform well on smaller targets such as ACE inhibitors and dopamine agonists.

This size-related phenomenon has been noted in previous studies. Indeed, the binary form of the Tanimoto coefficient has occasioned some recent criticism. Several previous studies[20−22] have investigated the distributions of values when a similarity coefficient is applied to a set of bit-string representations of chemical structures and related these to their effectiveness in areas such as dissimilarity-based subset selection. These studies agree that most of the coefficients have a preference for certain value ranges.

This is further illustrated in Figure 4, which shows the distribution of bit densities (UNITY 2D) for all compounds retrieved in the top 5% for searches for 5HT4 agonists using the Tanimoto, Russell/Rao, and Forbes single measures. The single measures have an obvious preference for selecting compounds in certain size ranges. The Forbes tends to select smaller compounds, the Russell/Rao selects larger compounds, and the Tanimoto lies somewhere in between.

Table 3 also shows that the best combination rarely relates to simply choosing the top performers in the merit tables, like Tables 1 and 2, and that the choice reflects the way in which the coefficients complement each other. Indeed, it was found that some of the coefficients which performed very poorly individually were found to be some of the best when used in combination. The Forbes is a particular example of this situation as it was the worst performer when applied to searches for 5HT4 agonists and benzodiazepine agonists (using UNITY 2D and BCI bit-strings) but appeared in many of the best combinations, always with the Russell/Rao.

Figure 5 shows the distribution of bit densities (UNITY 2D) for the active targets in the top 5% structures returned by the Forbes and Russell/Rao single measures when 5HT4 agonists were used as targets. The plots indicate the distributions for the targets which are retrieved by Forbes only, by Russell/Rao only, and those which are common to both coefficient searches. This illustrates the way in which the two coefficients complement each other, since the actives which are returned by the Forbes but are missed by the

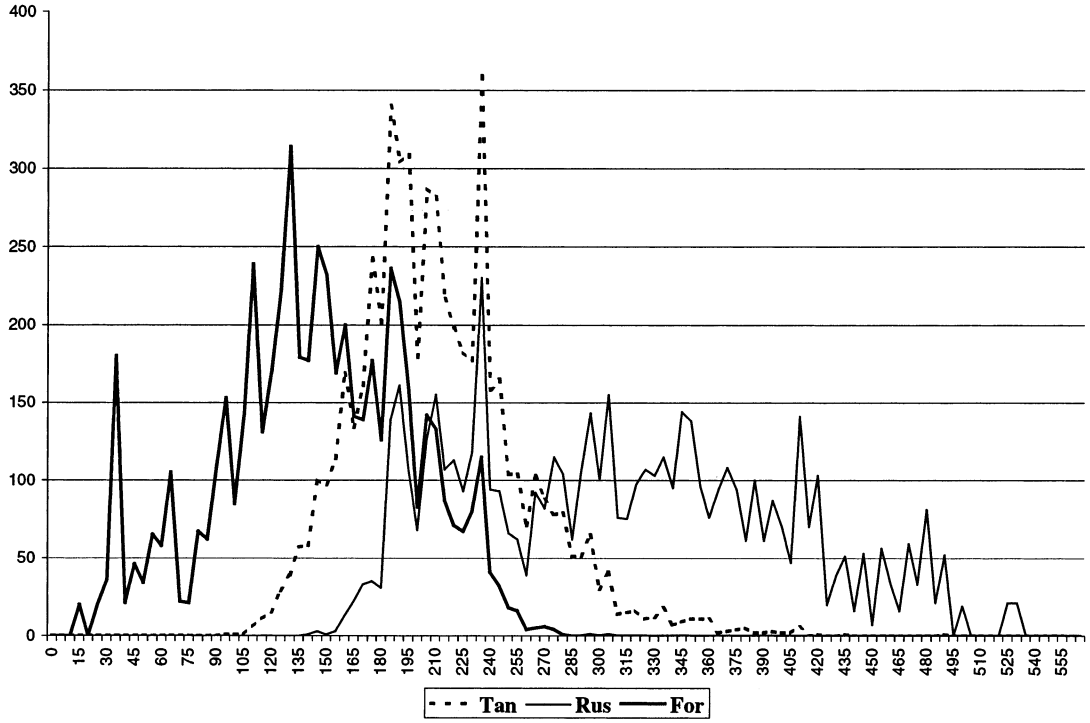

Figure 4. Distribution of number of bits set in top 5% structures obtained through similarity searching with 21 5HT4 agonist targets. Compounds are characterized by UNITY 2D bit-strings.

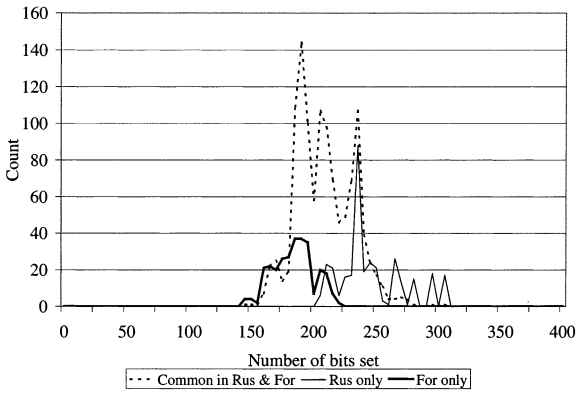

Figure 5. Bit set distribution of common and uncommon actives among the top 5% structures returned by the Russell/Rao and Forbes coefficients. Compounds are characterized by UNITY 2D bit-strings: (a) performance in 2-fusions, (b) performance in 3-fusions, and (c) performance in 4-fusions.

Russell/Rao tend to be smaller in size, while those returned by the Russell/Rao and missed by Forbes tend to be larger.

These results indicate that combining coefficients can improve search performances when compared with searches using single coefficients. However, there is no one combination that works well for all situations. The size of actives, the query itself, and the actives within the database seem to have a significant impact on the choice of coefficients. Initial results indicate that many combinations which include the Forbes are useful in searches for small actives and those which include the Russell/Rao are useful in searches for large actives. In many cases, the complementary effect of the Forbes and the Russell/Rao, possibly with one or two other coefficients, has resulted in combinations which perform well.

The optimum number of coefficients to combine seems to be between two and four. This is supported by Wang and Wang's study[23] on binding affinity estimation using multiple scoring functions. They demonstrated that consensus scoring improves hit-rates in virtual library screening simply because the mean value of repeated sampling tends to be closer to the true value, making it more robust and accurate compared to any single scoring procedure. The enhancement over use of single coefficients became less effective when more than four scoring functions were included in the consensus scoring procedure.

FURTHER EVALUATION STUDIES

In a further study, another 30 active classes, with each class having between 100 and 700 database compounds which exhibit the respective activity, were selected from the MDDR database. These were divided into three groups of 10 classes. For each of the three groups, all compounds which exhibited one of the 10 active classes, but which did not exhibit any of the other nine activities in the group, were placed in a subset. This was repeated for all 10 activity classes in the group, resulting in three subsets containing compounds which exhibited one, and only one, of the 10 activities in the group. In addition, several additional compounds, taken from every 20th compound in the MDDR database, were added to each subset. These were not added if they exhibited any of the 10 activities in the respective group. The three subsets were characterized by UNITY 2D bit-strings, Daylight fingerprints, and BCI bit-strings, respectively. Searches were carried out using 21 actives from each of the 30 activity classes as queries against their respective subset. This time, the performance measure of each search was taken as the number of actives in the top 400 nearest neighbors rather than the top 5%.

Table 4 shows the results for the best 2- to 4-coefficient fusions using UNITY 2D and Daylight characterizations (BCI results were similar). The active classes are ordered

Table 4. Performance of Best Single Coefficient(s), the Tanimoto Coefficient, and the Best Combinations for Selected Data Sets from the MDDR Dataset

target	av bits set	bestInd	num. act in best single	tan	best 2-fusion	num. act in best 2-fusion	best 3-fusion	num. act. in best 3-fusion	best 4-fusion	num. act. in best 4-fusion
Unity 2D (Average Bits Set = 221.84, Std. Dev. 70.78)										
enkephalinase inhibitor	163	For	53.67	44.90	For,Yul	52.86	SM,For,Den	52.43	SMFor,Yul,Den	51.48
prolylendopeptidase inhibitor	184	Den	129.57	124.19	Tan,For	131.81	Tan,SM,For	131.48	TanSM,Cos,For	131.76
dopamine (D2) antagonist	198	Den	57.57	52.38	For,Fos	58.76	Tan,For,Yul	58.76	For,Sim,Sti,Cos/Fos	62.42
vasolidator	206	Den	49.10	47.19	Bar,Yul	49.62	Bar,Pea,Yul	49.19	Bar,Pea,Yul,Den	49.19
thromboxane antagonist	209	Den	154.71	148.43	For,Cos/Fos	154.48	Tan,Ku2,For	155.76	Cos,For,Fos,Pea or Tan,Bar,For,Fos	155.62
excitatory amino acid inhibitor	209	Bar/Pea	44.62	43.52	Tan,For	44.43	Rus,SM,For	44.86	Rus,SM,For,Cos/Fos	45.24
potassium channel blocker	220	Cos	44.81	44.10	Rus,Bar/Den	47.10	Rus,Ku2,For	46.62	Rus,For,Fos,Sim	46.38
leukotriene D4 antagonist	224	Bar	129.48	128.14	Rus,For	131.67	Rus,SM,Sim	133.62	Tan,Rus,For,Sim	133.86
factor Xa inhibitor	240	Ku2	101.62	99.24	Ku2,Pea	102.2	Cos,Ku2,Pea	102.95	Rus,Ku2,For,Sim	103.29
endothelin ETA antagonist	272	Rut	38.76	37.57	Rus,Bar	41.90	Tan,Rus,Cos	41.24	Tan,Rus,Cos,Fos	40.71
Daylight (Average Bits Set = 287.02, Std. Dev. = 108.73)										
melatonin agonist	193	For	74.10	70.00	SM,Den	75.00	SM,Ku2,For	75.29	SM,For,Yul,Sti	75.33
adrenergic (alpha2) agonist	219	Bar	45.90	43.90	SM,Ku2	46.52	SM,Cos,For	46.71	SM,Bar,Cos,For	46.48
myocardial antiischemic	229	For	15.67	8.90	SM,For	24.24	SM,For,Yul	23.76	SM,Ku2,For,Fos	23.52
adrenoceptor (beta3) agonist	255	Pea or Sti	127.10	124.29	Fos,Pea or Cos,Sti	127.19	Rus,SM,For	129.00	Rus,SM,For, Sim	131.57
neutral endopeptidase inhibitor	255	Fos	93.57	92.52	Tan,Fos	93.81	Rus,Bar,For	94.14	Rus,Cos,Ku2,For	94.90
H+/K+ − ATPase inhibitor	276	Sti	150.24	148.29	Tan,Den	150.57	Rus,SM,Sim	150.95	Rus,SM,For,Sim	151.29
IL-8 inhibitor	281	Ku2	41.48	39.38	Sim,Pea	42.38	Tan,Sim,Yul	42.48	Ku2,Sim,Pea,Yul	42.38
xanthine	299	Cos/Fos	106.57	106.52	any fusion of Tan,Cos, Ku2,Fos	106.57	Rus,Bar,Sim	106.86	any fusion of Rus,Sim with best singles	106.81
phosphodietterase inhibitor	332	Tan	31.48	31.48	Rus,Pea	33.52	Rus,Ku2,Pea or	32.62	Tan,Rus,Cos,Fos	32.29
adenosine (A2) antagonist	348	Rus	32.10	29.00	Rus,Sim	33.38	Rus,Ku2,Sim	33.00	Rus,Ku2,Sim,Cos/Fos	32.19

Table 5. Performance of Best Single Coefficient(s), the Tanimoto Coefficient, and the Best Combinations for Selected Actives from the ID Alert Database.

target	av bits set	bestInd	num. act in best single	tan	best 2-fusion	num. act. in best 2-fusion	best 3-fusion	num. act. in best 3-fusion	best 4-fusion	num. act. in best 4-fusion
Unity 2D (Average BS 207, Std. Dev. = 71.11)										
PAF antagonist	167	For	2.84	2.13	SM,For	2.71	SMFor,Pea	2.68	SM,For,Yul,Den	2.64
TXA2 antagonist	184	For	2.23	1.62	For,Sim	2.12	For,Sim,Yul	2.11	For,Sim,Pea,Den	1.92
HMG CoA reductase inhibitor	204	Cos/Fos	2.03	2.00	Rus,Den	2.20	Rus,Bar,Pea	217	Rus,SM,Sim,Tan/ Cos/Bar/Fos/Ku2	2.1
angiotensin 11 antagonist	211	Cos/Fos	5.83	5.81	Rus,Pea	6.11	Tan,Rus,Cos/Fos	6.10	Rus,Ku2,Fos,Pea	6.01
tyrosine kinase inhibitor	218	Rus	2.39	2.00	Rus,Sim	2.30	Tan,Rus,Bar	2.13	Rus,Ku2,Yul,Cos/Fos	2.0
ACAT inhibitor	232	Rus	3.35	2.37	Rus,Sim	2.96	Rus,Ku2,Sim	2.89	Rus,Ku2,Sim,Cos/Fos	2.83
cephalosporin	245	Rus	2.50	2.10	Rus,Sim	2.80	Rus,Sim,Yul	2.65	Yul,Cos/Fos/Ku2	2.60
Daylight (Average BS = 269, Std. Dev. = 111.59)										
antiarrhythmic agent	163	For	1.83	1.33	SM,Sim	2.00	SM/Sim,Fos/Sti	2.00	SM,Fos,Sim,Sti	2.00
fibrinogen antagonist	213	For	2.16	1.53	SM,Sim	2.21	SM,Sim,Fos/Sti	2.21	SM,Fos,Sim,Sti	2.21
calcium channel blocker	252	Den	2.78	2.46	Tan,SM	2.93	Tan,SM,Fos/Sti	2.93	Tan,SM,Sim,Sti	2.93
potassium channel blocker	275	Tan	2.72	2.72	Tan,Fos/Sti	2.72	Tan,Fos,Sti	2.72	Tan,Cos,Fos,Sti or Rus,Cos,Sim,Yul	2.66
estrogen antagonist	284	Tan/Bar/Pea/ Sti/Den	1.18	1.18	Fos,Sti	2.00	Rus,For,Fos/Sti	1.45	Rus,For,Fos,Sti	1.45
acetylcholinesterase inhibitor	307	Bar/Pea/Sti	1.78	1.67	Fos,Sti	2.00	Rus,For,Sim	1.94	Rus,For,Sim,Cos/ Ku2/Fos/Sti	1.94
PDE inhibitor	322	Rus	1.95	1.21	Fos,Sti	2.00	Rus,Fos,Sti	1.95	Rus,Fos,Sim,Sti	1.63
muscarinic MI agonist	337	Rus	1.17	1.00	Rus,Fos/ Sim/Sti	1.17	Rus,Sim,Ku2/Fos or Rus,Sti, Fos/Sim	1.17	Rus,Sim & multiple fusions of Ku2,Fos,Sti	2.33

by bit density and illustrate two points. First, that, in most cases, the best combination performs better than the best single coefficient. Second, it is apparent that the size (or bit density) affects the choice of coefficients. Smaller actives, like Enkephalinase inhibitors or Melatonin agonists, show better performance where the Forbes or Simple Match are in combination, whereas larger actives, like Endothelin ETA antagonists or Adenosine (A2) antagonists, show better performance when the Russell/Rao is in a combination, usually with the Tanimoto, Fossum, and/or Cosine. Many of the mid-sized actives show good performance with the Russell/Rao and the Forbes or Simple Match in combination due to the complementary effect discussed above.

Table 4 also shows the result for the Tanimoto coefficient which, in nearly all cases, is out-performed by another single coefficient and is out-performed by a combination every time.

A further study used three sets of seven activity classes taken from the IDAlert database, selected by ordering all

activity classes by their average bit density and, for each set, selecting seven classes which covered the entire range of bit density for the entire database. For each set, one of the three characterizations (UNITY 2D bit-strings, Daylight fingerprints and BCI bit-strings) was used, and all actives from the seven active classes were used as queries against the whole database. Again the performance measure was the number of actives in the top 400 nearest neighbors. The results, exemplified in Table 5 by UNITY 2D bit-strings and Daylight fingerprints, were similar to the MDDR study, with the Forbes and Simple Match being most suitable in combinations for searches involving small actives, and the Russell/Rao being favored in combinations for searches involving large actives.

DISCUSSION OF FURTHER EVALUATION STUDIES

The results indicate that a good combination of the right number of coefficients can improve retrieval performances over single measures and, in particular, over the industry standard measure, the Tanimoto coefficient. However, there is no one combination which performs equally well on all active classes. Indeed, a best-performing combination for one type of active is very often a poor performer for another. Our dilemma is that, to know which combination to use, we need to try all possible combinations, and single measures, for comparison. There is no combination or single measure which, on average across all types of activity, performs to a high standard. A few would be preferable for use on all active classes, including the Tanimoto single measure, as they would give a better than average performance in most cases.

We now know that the dominating factor determining the choice of coefficient is the size of the active compounds in the class of interest, and we know which are the preferred coefficients for the various size ranges. We have also reduced the pool of coefficients to select from, as certain coefficients, whether single or in combination, appear repeatedly as best performers. One would definitely include the Russell/Rao, Forbes, and Simple Match in this pool and would probably add the Tanimoto, Baroni-Urbani/Buser, Kulczynski (2), Fossum, and Cosine. These are also good performers when used as single measures. There is a case for using all of these in combination, with the addition of a weighting scheme to increase the effect of the more desirable coefficients. Again, this requires prior knowledge to determine the correct weighting for the characteristics of the active class of interest.

CONCLUSIONS

In this paper we have presented the results of several studies into data fusion of similarity coefficients for similarity searches of a selection of databases. The results indicate that combining coefficients does improve the performance of similarity searches when compared with the use of single measures, in particular the industry standard Tanimoto measure. The optimum number of coefficients to use in combination tends to be between two and four with the improvement diminishing at five or more coefficients.

The performance of the coefficient combinations depends on the distribution of bits set in the bit-strings representing the compounds involved and, in particular, their bit densities. For smaller compounds performance is increased when using a combination containing the Forbes and Simple Match, whereas for larger compounds performance is increased when using a combination containing the Russell/Rao. Combinations containing coefficients which perform well when used as single measures tend to perform well across most of the size ranges.

We can conclude that, although combinations are generally better alternatives to single measures, the practicality of their use remains questionable as no one combination showed consistently high performance across all types of activity. The Tanimoto coefficient is a good single measure as no other combinations gave consistent improvement over its use.

ACKNOWLEDGMENT

We thank The National Cancer Institute, Molecular Design Ltd., and Current Drugs for the provision of the databases, Tripos Inc., Barnard Chemical Information, and Daylight for the provision of software, and The Wolfson Foundation and The Royal Society for hardware resources. We also thank Prof. Chang-Yu Hu and the KC Wong Education Foundation for much of the initial evaluation. Naomie Salim was funded by the Government of Malaysia. The Krebs Institute for Biomolecular Research is a designated biomolecular sciences center of the Biotechnology and Biological Sciences Research Council.

NOTE ADDED AFTER ASAP POSTING

This article was released ASAP on 12/20/2002 with an incorrect caption for Figure 2. The correct version was posted 12/23/2002.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572−584.
(3) Wipke, W. T.; Rogers, D. J. *J. Chem. Inf. Comput. Sci*. **1984**, *24*, 71−81.
(4) Bohm, H.-J.; Schneider, G., Eds. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, 2000.
(5) *Molecular Diversity in Drug Design*; Dean, P. M.; Lewis, R. A., Eds.; Kluwer: Amsterdam, 1999.
(6) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 65−84.
(7) Brown, R. D. *Perspect. Drug Discov. Des.* **1997**, *7/8*, 31−50.
(8) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(9) Holliday, J. D.; Hu, C.-Y.; Willett, P. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155−166.
(10) The NCI AIDS database is available from NCI/NIH Development Therapeutics Programme at URL http://dtp.nci.nih.gov/.
(11) The IDAlert database is available from Current Drugs Limited at URL http://www.current-drugs.com/.
(12) UNITY is available from Tripos Inc. at URL http://www.tripos.com.
(13) The MDL Drug Data Report database is available from Molecular Design Limited at URL http://www.mdli.com.
(14) Barnard Chemical Information Ltd. is at URL http://www.bci.gb.com.
(15) Daylight Chemical Information Systems Inc. is at URL http://www.daylight.com.
(16) Ellis, D.; Furner-Hines, J.; Willett, P. *Perspect. Inf. Manag*. **1994**, *3*, 128−149.
(17) Mojena, R. *Computer J*. **1977**, *20*, 359.
(18) Willett, P.; Winterman, V.; Bawden, D. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(19) Ginn, C. M. R.; Willett, P.; Bradshaw, J. *Perspect. Drug. Discov. Des*. **2000**, *20*, 1−6.
(20) Godden, J. W.; Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 163−166.
(21) Flower, D. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(22) Dixon, S. L.; Koehler, R. T. *J. Med. Chem*. **1999**, *42*, 2887−2900.
(23) Wang, R.; Wang, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

CI025596J