Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients

Martin Whittle,* Valerie J. Gillet, and Peter Willett

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

Alexander Alex and Jens Loesel

Pfizer Global Research and Development, Pfizer Limited, Ramsgate Road, Sandwich, Kent CT13 9NJ, U.K.

Received April 20, 2004

This paper evaluates the effectiveness of various similarity coefficients for 2D similarity searching when multiple bioactive target structures are available. Similarity searches using several different activity classes within the MDL Drug Data Report and the Dictionary of Natural Products databases are performed using BCI 2D fingerprints. Using data fusion techniques to combine the resulting nearest neighbor lists we obtain group recall results which, in many cases, are a considerable improvement on standard average recall values obtained for individual structures. It is shown that the degree of improvement can be related to the structural diversity of the activity class that is searched for, the best results being found for the most diverse groups. The group recall of active compounds using subsets of the class is also investigated: for highly self-similar activity classes, the group recall improvement saturates well before the full activity class size is reached. A rough correlation is found between the relative improvement using the group recall and the square of the number of unique compounds available in all of the merged lists. The Tanimoto coefficient is found unambiguously to be the best coefficient to use for the recovery of active compounds using multiple targets. Furthermore, when using the Tanimoto coefficient, the "MAX" fusion rule is found to be more effective than the "SUM" rule for the combination of similarity searches from multiple targets. The use of group recall can lead to improved enrichment in database searches and virtual screening.

INTRODUCTION

The search for new lead compounds in the pharmaceutical and agrochemical industries increasingly makes use of virtual screening of corporate, public, and virtual databases. Among the methods available, similarity searching is a cheap and widely used method to distill a pool of potentially interesting compounds from a large database. Underlying this method is the idea, often called the similar property principle,[1] that compounds with similar structures tend to express similar biological activity. If the exact nature of the structure−activity relationship is understood, then it may be possible to use alternative methods such as docking algorithms to find more specific suitable candidates. Typically, however, it may not be known how an active compound works and the selection of candidates for screening must be based on knowledge of the active structure itself. In these cases, algorithms can be used to search libraries for compounds that are structurally similar to a known active molecule, the *target structure* (sometimes called the query), and among these we would hope to find other compounds with comparable activity.

There are many ways of quantifying the similarity of two structures and the important point is to use a measure that maximizes the similarity in biological activity. Working on the hypothesis that different similarity measures probe different aspects of the underlying structure−activity relationship, several attempts have been made to use the techniques of data fusion[2] to improve the results of individual measures.[3−5] These efforts have met with mixed success, a major difficulty being the prediction of which particular combination(s) of similarity measures will lead to significant enhancement in search performance in a particular case.[6] However, if a group of active molecules is available, rather than just a single target structure, then they themselves map out a hypervolume in the multidimensional descriptor space of the chosen structure representation and in this way may be able to delineate the boundaries of available structures for the particular activity. The results of a search using a single structure may produce a set of nearest neighbors that spans only part of this space, but by combining the nearest neighbors from all available actives in a suitable way we may expect to cover more or even the whole of the space.

Data fusion has hitherto been applied to the combination of results from different similarity measures. In these studies, two or more nearest neighbor lists for each single active target structure, obtained by different similarity coefficients or different structure representations, are combined according to one of several fusion rules. In this paper we consider a

* Corresponding author e-mail: m.whittle@sheffield.ac.uk.

10.1021/ci049867x CCC: $27.50
Published on Web 09/02/2004

very different approach, viz. the combination of nearest neighbor lists from *different* active target structures for a single similarity measure. In the sense that the results from different target structures are being combined it may be argued that this procedure is not strictly data fusion, which classically combines the results of different sensors. Nevertheless, in the remote sensing community,[7] for example, different spectral channels of the same sensor are now frequently considered as distinct sources. Here, we have adopted an essentially analogous idea in the context of chemical similarity searching.

METHODS

Similarity searching begins with the identification of a known bioactive molecule, the target structure, which could be a published or commercially available compound or a hit from a high-throughput screening experiment. The structure of this molecule is then compared with each of the *comparison structures* in a database using an objective measure of similarity. The target and database may comprise 2D or 3D chemical structures, which are first rendered as a machine-readable *representation* that describes some structural features of the molecules under scrutiny. Weighting or standardization may then be applied before a *similarity coefficient* is used to quantify the degree of resemblance between the chosen representations of the target structure and each of the database structures.[8] The database molecules are then ranked in decreasing order of the calculated similarity values, with the top-ranked molecules, the *nearest neighbors*, being those that bear the closest structural similarity to the target structure. In a database of $\sim 10^5-10^6$ compounds typically $\sim 10^3$ of these neighbors might be chosen for subsequent investigation.

The effectiveness of the search depends on both the representation and similarity coefficient used, which together comprise the overall similarity measure. In this work we have compared the results from several similarity coefficients using 2D fingerprints from Barnard Chemical Information (BCI)[9] exclusively as the representation. These were calculated using the standard BCI fragment dictionary, which yields bit-strings containing 1052 bits to represent each of the molecules in a database that is to be searched. When such fragment-based bit-strings are used as the representation, the similarity coefficient between two structures is a straightforward function of a simple bit count. For bit-strings of length N, we will suppose that a bits are set in the string for a target compound A, b bits are set for a comparison molecule B, and c bits are set common to both strings. From these basic definitions all of the well-known similarity coefficients $S(A,B)$ can be calculated. On the basis of previous experience[10] a range of these have been used in this study, some of which are expected to give closely comparable results and some of which are quite different in character. The relevant expressions are collected in Table 1.

Scaling for Fusion. In our previous studies on data fusion[3−5] we have used a fusion rule based on similarity ranks. As noted by Ginn et al.[3] and by Wang and Wang,[11] this not only provides a simple form of standardization for combining measures that result in markedly different scores or differences of scores but also implies some loss of the information content inherent in a similarity ranking. Here, we have sought to avoid this drawback by combining similarity values directly. In a similarity search that produces r nearest neighbors the rth similarity value will normally differ between targets. For this reason it is useful to work with scaled versions that ensure comparable ranges for the results of each target structure. Ng and Kantor[6] have previously discussed the use of linear Range Scaling[12] for data fusion, and we have used this idea here. For each unscaled similarity result $S(A,B)$ between target A and comparison structure B a rescaled value $S^*(A,B)$ is obtained using

$$S^*(A,B) = \frac{S(A,B) - S_{\min}(A)}{S_{\max}(A) - S_{\min}(A)} \qquad (1)$$

where $S_{\max}(A)$ and $S_{\min}(A)$ are the maximum and minimum values in the ranked list for target A. This transformation maps the original values onto the range 0−1, but since it is a simple linear scaling there can be no difference between the rank positions of comparison structures scored using the scaled or unscaled version of the result. However, the *relative* rank of comparison structures in lists for different target structures may be very different, and this could potentially affect the results of data fusion; note also that even if $S(A,B) = S(B,A)$ for a particular coefficient, the equivalent equality for scaled coefficients does not necessarily hold. In practice, however, this scheme has a number of advantages: it puts all similarity measures on an equal footing; it automatically gives equal weight to different targets, while being amenable to alternative weighting if desired; and it makes the fusion of incomplete lists straightforward, since a score of zero can be associated with absent entries.

Table 1. Similarity Coefficients Used in This Work[a]

coefficient	abbreviation	expression
Tanimoto	T	$S_T = c/(a + b - c)$
Modified Tanimoto	MT	$S_{MT} = S_T(2 - \rho_0)/3 + S_{T0}(1 + \rho_0)/3$
Cosine	C	$S_C = c/\sqrt{ab}$
Squared Euclidean	E	$S_E = (a + b - 2c)/N$
Kulczynski(2)	K	$S_K = (1/2)[c/a + c/b]$
Baroni-Urbani	BU	$S_{BU} = (\sqrt{cd} + c)/(\sqrt{cd} + a + b - c)$
Pearson	P	$S_P = (Nc - ab)/\sqrt{Nab(N - b)(N - a)}$
Russell-Rao	R	$S_R = c/N$
Forbes	F	$S_F = cN/(ab)$
Simpson	SI	$S_{SI} = c/\min(a,b)$
Yule	Y	$S_Y = (Nc - ab)/(cd + (a - c)(b - c))$

[a] The definitions apply for the combination of bit-strings of length N where a bits are set in the target compound string, b bits are set in the comparison string, c bits are set common to both strings, and d bits are set in neither string. The expression for the Modified Tanimoto coefficient includes the average bit density ρ_0 and $S_{T0} = d/(N - c)$ the Tanimoto coefficient for absent features.

Cumulative Recall. The effectiveness of the similarity searches considered here is described by the *cumulative recall*.[13] Suppose that the target molecule, i, is one of n known active molecules and that we examine the nearest neighbors recovered in a similarity search down to rank r. Among these nearest neighbors, the search might recover $a_i(r) \leq (n - 1)$ members of the activity class and the cumulative recall $R_i(r)$ at rank r is given by the ratio

$$R_i(r) = \frac{a_i(r)}{n - 1} \qquad (2)$$

Table 2. Compound Classes Used in This Study[a]

activity class	abreviation	class	n	μ	σ
Hhunity/DNP 9.1	Hhunity	mixed, active, natural products	50	0.265	0.221
Wound healing agents	WoundHeal	dermatological agents	83	0.334	0.189
Dopamine autoreceptor agonists	Dop_Ag	antipsychotics	170	0.391	0.172
HIV1 protease inhibitors	HIVPI	agents for AIDS	93	0.389	0.155
Reverse transcriptase inhibitors	RTI	agents for AIDS	91	0.289	0.139
Penicillins	Pen	β-lactam antibiotics	89	0.525	0.171
Acetyl-cholinesterase inhibitors	Achinb	agents for cognition disorders	94	0.377	0.144

[a] The sets used contained a random selection of active molecules from the complete activity classes. For these, n is the number of target structures chosen, μ is the mean, and σ is the standard deviation for the intraclass similarities obtained using the Tanimoto coefficient.

Using each of the actives in turn we can then obtain an average recall value as

$$R_{av}(r) = \frac{1}{n}\sum_{i=1}^{n} R_i(r) = \frac{1}{n}\sum_{i=1}^{n} \frac{a_i(r)}{n-1} \tag{3}$$

The average recall, eq 3, is just related to the expected or average number of actives that would be recovered by any single randomly chosen active target. These values are, to an extent, dependent upon the representation and similarity coefficient used but they are also strongly dependent on the particular activity class under consideration. If the activity class is highly self-similar under the metric chosen, then, unsurprisingly, relatively high values of the average recall are returned.

Active Compound Sets. Seven sets of active compounds have been chosen for this study: six from the MDL Drug Data Report (MDDR)[14] and, as a comparative link with an earlier study,[4] one set of mixed pharmacological compounds from the Dictionary of Natural Products version 9.1 (DNP).[15] The number of chosen compounds with valid fingerprints used to calculate the recall is noted in Table 2.

Searches using these target sets were performed, as appropriate, on the 102443 compounds from the MDDR database and on the 107052 compounds from the DNP database that remained after filtering to remove blank records and duplicate structures. A guide to the self-similarity of the active sets is provided by matching each compound with every other in a group, calculating similarities using the Tanimoto coefficient and computing the mean, μ, and standard deviation, σ, of the results. These values are also given in Table 2, where it will be seen that the penicillins are the most self-similar group and the Hhunity set from the DNP is the least self-similar.

Fusion of Lists for Multiple Actives. From the average recall, eq 3, the average number of actives recovered by using any single active as the target is given by $a_{av}(r) = (n-1)R_{av}(r)$. If $m \leq n$ compounds, which we will call the primary targets, are chosen from the active set we will obtain m lists each of length r and the total number of records in all the lists is mr. By concatenating all of the lists a total number of actives, a_{tot}, may be recovered (obviously excluding the targets heading each list), which could include some of the $n - m$ secondary targets that have not been chosen to perform the search but are nevertheless known to be active. If each of the targets recovers almost the same sets of active compounds in its nearest neighbor lists, then we would expect that $a_{tot} \approx a_{av}$, but if each target recovers significantly different sets of actives, then all of the actives might in principle be retrieved in the concatenated list so that $a_{tot} \approx n$; i.e, in practice, $a_{av} \leq a_{tot} \leq n$. From the number of actives recovered in the concatenated list we can then define a total recall that depends on the rank and the size of the activity class

$$R_{tot}(r,m,n) = \frac{a_{tot}(r,m,n)}{n} \tag{4}$$

In practice this would mean screening the full set of mr compounds, rather than the r compounds from a single list. This may be feasible for small numbers of targets, but it unlikely to be a practical proposition for large numbers of primary targets and furthermore it is not a method that can be fairly compared with the results for single actives. Fortunately, it is possible to apply a data fusion algorithm to the set of individual target lists and thus produce a single list of comparable length to the originals while still recovering some of this total number of actives. Essentially, for all m target compounds i and comparison structures j that are found in the ranked lists of length r we compute the fused result $S_{FUS}(j)$ from a fusion rule F and the scaled similarities, $S^*(i,j)$ for a particular coefficient, as obtained from eq 1

$$S_{FUS}(j) = F_{i=1}^{m}[S^*(i,j)] \tag{5}$$

If structure j is not found in one of the lists the scaled similarity is assumed to be zero. The fusion rule, F, indicates the method used to combine the similarities obtained for the m targets. The SUM rule has been used for most of the work reported here, and in this case a summation sign Σ would replace F. However, we have also obtained some results using the MAX rule in which case F operates on the set of similarities to choose the maximum value for each j. Other rules are also possible and have been discussed elsewhere.[3] The structures j are then ranked according to the returned values of the combined similarity. The method has much in common with the schemes presented by Wilton et al.[16] However, in our case the similarity values $S^*(i,j)$ are first ranked and truncated at a specified value of r and then scaled according to eq 1. This ensures that all of the comparison structures considered for fusion have a comparable level of similarity to the target set and reduces the number of values that are processed. The rescaling also alters the relative scores of compounds in the set of individual lists and thus changes the resulting final scores. In particular, the scaling will help to reduce any bias due to clusters of very similar molecules, all of which will return high, unscaled scores. By choosing only those r test-set and training-set pairs that have maximum similarity the initial ranking of results performed by the present fusion algorithm is automatically assured. Thus, apart from the rescaling used in this work, the method used by

Wilton et al.[16] is expected to be identical to the data fusion algorithm when F stands for "choose the maximum value". The calculation of group recall values is discussed further below.

Group Recall. Although some of the nearest neighbors for two target lists may be common to both, there may also be some that are different. This applies equally to those nearest neighbors identified as members of the active set (i.e. contributing to the recall) and those that are not. A fusion algorithm that merges two lists of length r by combining the similarities of common compounds will therefore result in a fused list of length Z_2 that will usually be longer than the original length r but shorter than $2r$; hence $r \leq Z_2(r) \leq 2r$ where the suffix indicates the order of fusion. More generally, the combination of all m lists will generate a final list of length $Z_m(r)$, where $r < Z_m(r) < mr$. The lengthening of the fused list as the number merged increases depends on the relative composition of the individual lists and leads naturally to the definition of a useful measure that we call the overall *disparity*, D, which is just the ratio of the total length of the fused list to the maximum possible length

$$D(r) = \frac{Z_m(r)}{mr} \tag{6}$$

If all the lists are completely different, then this has a value of 1.0, while if all the lists are the same, then it has a value of $1/m$.

The next step in the fusion algorithm is to take the list of length $Z_m(r)$, rank by the fused similarity values, eq 5, and truncate the resulting list at rank r for comparison with the original lists. The final merged list does not correspond to any particular target, and a known active molecule may not necessarily occupy the top position but it will nevertheless contain members of the active set so that a recall value can be defined. The number recovered depends on the rank, number of primary targets used, and the total number of active targets searched for. Thus we write the number recovered as $a_G(r,m,n)$ and the group recall as the ratio

$$R_G(r,m,n) = \frac{a_G(r,m,n)}{n} \tag{7}$$

Note that, when combining lists by summation, the self-similar value of the target compound is ignored. This is so that if none of the original lists contain compounds other than the target headers the value of $R_G(r,m,n)$ is zero. Many of our results have used all members of the active set in the search in which case $m = n$. In much of what follows the parameters are implicit, and, for example, R_G is used in place of $R_G(r,m,n)$ for simplicity.

In another variation, it is possible to sum the rank positions of the individual lists rather than the similarity values and to obtain a rank group recall value $R_{GR}(r,m,n)$ by analogy with eq 7. Results for all of these recall quantities — R_{av} (eq 3), R_G (eq 7), and R_{GR} — are given in Table 3a–g for the seven groups of compounds identified in Table 2. These tables also contain values for ΔR, the *group recall fractional improvement index*, that is defined and discussed in the Results section.

RESULTS

We start by discussing group recall results obtained by using all available actives as targets, i.e., by using $m = n$ in eq 5. Unless otherwise specified, fusion has been performed using the SUM rule. Virtual screening experiments are normally reported using a cumulative recall plot such as that shown in Figure 1, which relates the number of active structures retrieved to the total number of structures (both active and nonactive) retrieved.[4,13]

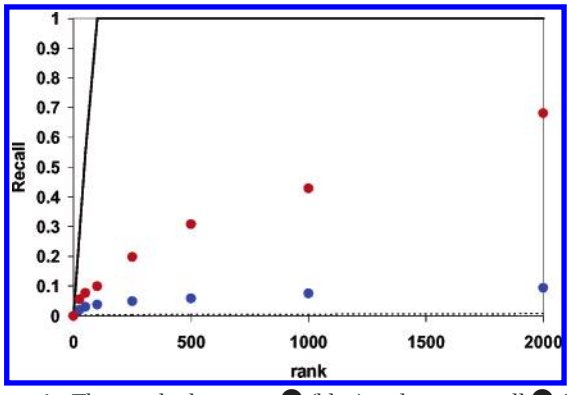

Figure 1. The standard average ● (blue) and group recall ● (red) obtained for the activity class of reverse transcriptase inhibitors (RTI) plotted against the rank. Also shown are the following: — the theoretical maximum and ··· the random expectation lines.

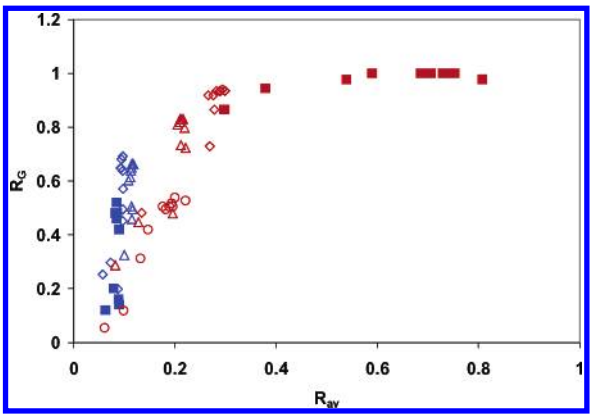

Figure 2. Group recall R_G vs the average recall R_{av} categorized by activity class. The results were obtained at rank 2000. For each group the results for all 11 similarity coefficients are included. Groups with high self-similarity (Table 3): ■ (red) Pen; ◊ (red) Dop_Ag; ○ (red) HIVPI; △ (red) Achinbs; groups with relatively low self-similarity: ◊ (blue) RTI; △ (blue) WoundHeal; ■ (blue) Hhunity.

Figure 1 makes a direct comparison between group recall, eq 7, and the average recall result, eq 3, obtained for a single target, as measured for the single activity class of reverse transcriptase inhibitors (RTI) using the Tanimoto coefficient. The results are plotted against the rank at which they were obtained along with lines representing the theoretical maximum and random expectations. There is clearly a significant improvement achieved by using group fusion, with the degree of improvement increasing with rank. The degree of improvement was found to vary over both the activity class and similarity coefficient used. Figure 2 compares the group recall result with the average results for all activity classes.

In almost all cases the group recall R_G values obtained by fusing results from the whole target set are a significant improvement on the standard average recall R_{av}. In this figure, results are also coarsely color-coded according to the mean similarities of the groups given in Table 2. This reveals, as might be expected, that a high self-similarity of the group is reflected in relatively high values of both group and

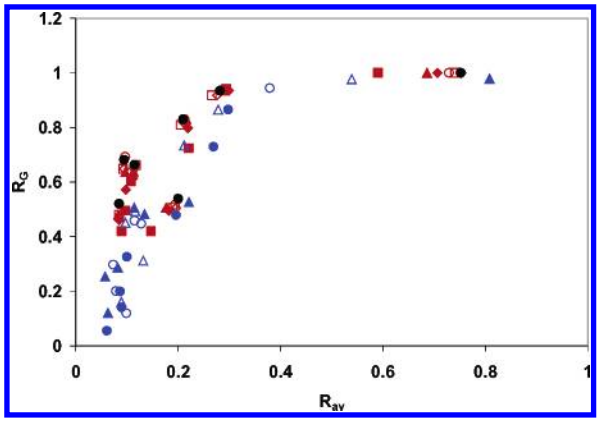

Figure 3. Group recall R_G vs the average recall R_{av} categorized by similarity coefficient. The results were obtained at rank 2000. For each coefficient the results of all seven activity groups are included. "Regular" coefficients: ● (black) Tanimoto; ○ (red) Modified Tanimoto; ◇ (red) Cosine; □ (red) Kulczynski(2); ▲ (red) Pearson; ■ (red) Sq. Euclidean; ◆ (red) Baroni-Urbani; "irregular" coefficients: (blue) ▲ Forbes; ● (blue) Rusell-Rao; ○ (blue) Simpson; △ (blue) Yule.

average recall. However, although a low self-similarity generally means a low average recall, it will be seen that the group recall approach can offer a significant improvement.

The same results are redisplayed in Figure 3, this time categorized by similarity coefficient. A recent mathematical and graphical analysis[10] of the behavior of similarity coefficients focused on details of the dependence of similarity values on molecular weight through the relative bit-density. Four of the coefficients studied in ref 10 were found to have unusual behavior, these being the Russell-Rao, Forbes, Simpson, and Yule coefficients. These "irregular" coefficients have frequently given either improved or degraded results in the past compared to the other "regular" similarity coefficients,[4] and indeed we find here that within each activity group the results for these four coefficients show relatively small improvement. This is emphasized in Figure 3, where results from these coefficients are distinguished by color from those derived using the well-behaved coefficients.

This observation is reinforced by ranking the similarity coefficients by group recall value for each activity class, using an average value for any ties and then ranking the coefficients according to the overall average taken over all activities. These results are given in Table 4a where the four irregular coefficients take the lowest positions. From a practical point of view it is also clear from the values in Table 4a that the simple Tanimoto coefficient is unequivocally the best choice of coefficient for group recall.

The degree of improvement over the use of a single active can be quantified by defining a fractional improvement index relative to the average recall value

$$\Delta R = \frac{R_G - R_{av}}{R_{av}} \qquad (8)$$

Values of ΔR are included in the right-hand column of Table 3 and plotted against the average recall in Figure 4, where the color coding again distinguishes between the results obtained using regular and irregular coefficients.

This figure also shows the maximum possible value of ΔR, obtained when $R_G = 1.0$ in eq 8, and it is clear that this value is achieved by some of the results (the penicillins Table 3f) at high values of R_{av}. It is also clear that as R_{av} approaches 1.0 (the theoretical maximum) ΔR decreases, and it becomes increasingly difficult to achieve significant improvement using group recall. At lower values of R_{av} our results for ΔR fail to reach this maximum but nevertheless show a significant improvement. Indeed, the best improvement is obtained for results with an average recall of about 0.1, which is the lowest investigated. Referring to Figure 2, we see that this corresponds to activity classes with a low mean self-similarity (the blue-coded points), i.e., the relative improvement resulting from fusion is greatest with heterogeneous sets of actives where individual searches perform poorly. Results with high values of R_{av} to the right of Figure 4 show only small improvements using group recall, and these correspond to groups with a high self-similarity such as the penicillins. For example, using the Tanimoto coefficient the values of ΔR increase in the order Pen < HIVPI < Dop_Ag < Achinb < WoundHeal < Hhunity < RTI, an order that is very similar to the order of decreasing self-similarity detailed in Table 2. In fact, a plot of ΔR against μ for these results gives quite a reasonable linear correlation ($R^2 = 0.85$) with gradient -22.72 and intercept 11.68. The overall average rank positions for the similarity coefficients obtained using the values of improvement index, ΔR, for each activity class are given in Table 4b. Notwithstanding ties, the resulting sequence generally agrees with that of Table 4a and reinforces our conclusions from those results. However, the table reveals that the Russell-Rao and Forbes coefficients yield the best improvement for the penicillins and dopamine agonists, respectively. The reason for this large improvement factor is that the average recall values for these results are particularly low within their respective classes (0.134 for Forbes/dopamine agonist as against a mean of 0.281 for the other coefficients and 0.298 for Russell-Rao/penicillin as against a mean of 0.668 for the other coefficients). Notably, according to Table 2, these activity classes are the most internally similar that we have studied.

The group recall by fused rank results, R_{GR} (Table 3), are almost always lower than the results obtained by summing similarities. Some loss of information occurs when results are ranked, particularly if ties occur, but the difference is nevertheless surprisingly significant. As a test of the technique, the ranks from some of the results were rescaled to give a score between 0 and 1, with the top rank taking the high score. These scores were then used in the routine for fusion by summation of similarities, and values identical to R_{GR} were obtained. We hence believe that the use of fused similarity is far preferable to the use of fused ranks.

Our study has focused on the use of the SUM fusion rule since this has provided good results in the past[3,5] for the combination of ranked results. As mentioned earlier, another popular combination rule is the MAX rule. When using this rule the fusion algorithm finds the r target and database pairs that have the maximum similarity by picking them from the m lists of r molecules deemed most similar to the targets by each measure. Group recall results using the MAX fusion rule with the scaled Tanimoto similarities are given in Table 5 alongside those obtained using the SUM rule. In all cases, the MAX rule yields equivalent or improved results. In one

Table 3. Similarity Results by Activity Class[a]

similarity coefficient	D	R_{av}	R_G	R_{GR}	ΔR	similarity coefficient	D	R_{av}	R_G	R_{GR}	ΔR
					a. Hhunity						
T	0.372	0.085	0.520	0.140	5.118	E	0.335	0.090	0.420	0.120	3.667
MT	0.369	0.085	0.520	0.140	5.118	SI	0.336	0.079	0.200	0.220	1.532
P	0.362	0.083	0.480	0.140	4.783	F	0.283	0.063	0.120	0.100	0.905
K	0.370	0.085	0.480	0.140	4.647	Y	0.365	0.089	0.160	0.160	0.798
C	0.372	0.086	0.480	0.140	4.581	R	0.251	0.090	0.140	0.120	0.556
BU	0.360	0.085	0.460	0.160	4.412						
					b. WoundHeal						
T	0.253	0.115	0.663	0.470	4.765	E	0.210	0.108	0.602	0.409	4.574
MT	0.250	0.114	0.651	0.482	4.711	SI	0.226	0.115	0.458	0.277	2.983
P	0.247	0.113	0.639	0.470	4.655	F	0.186	0.114	0.506	0.241	3.439
K	0.225	0.118	0.662	0.558	4.610	Y	0.241	0.116	0.494	0.458	3.259
C	0.253	0.116	0.663	0.458	4.716	R	0.161	0.100	0.325	0.096	2.250
BU	0.245	0.112	0.614	0.482	4.482						
					c. Dop_Ag						
T	0.112	0.282	0.935	0.724	2.316	E	0.084	0.294	0.941	0.706	2.201
MT	0.109	0.289	0.935	0.724	2.235	SI	0.064	0.269	0.729	0.518	1.710
P	0.109	0.290	0.935	0.759	2.224	F	0.085	0.134	0.482	0.400	2.597
K	0.120	0.266	0.918	0.729	2.451	Y	0.117	0.278	0.865	0.782	2.112
C	0.115	0.276	0.918	0.729	2.326	R	0.064	0.269	0.729	0.635	1.710
BU	0.104	0.299	0.935	0.765	2.127						
					d. HIVPI						
T	0.188	0.200	0.538	0.419	1.690	E	0.188	0.147	0.419	0.269	1.850
MT	0.190	0.194	0.516	0.430	1.660	SI	0.171	0.099	0.118	0.108	0.192
P	0.193	0.176	0.505	0.323	1.869	F	0.112	0.221	0.527	0.419	1.385
K	0.191	0.189	0.505	0.387	1.672	Y	0.188	0.132	0.312	0.226	1.364
C	0.189	0.196	0.505	0.430	1.577	R	0.140	0.061	0.054	0.054	−0.115
BU	0.193	0.182	0.495	0.333	1.720						
					e. RTI						
T	0.297	0.094	0.681	0.407	6.214	E	0.243	0.097	0.495	0.308	4.093
MT	0.294	0.097	0.692	0.407	6.156	SI	0.247	0.073	0.297	0.198	3.063
P	0.289	0.097	0.637	0.407	5.553	F	0.184	0.057	0.253	0.055	3.408
K	0.296	0.093	0.648	0.418	5.983	Y	0.270	0.097	0.451	0.330	3.649
C	0.300	0.094	0.648	0.418	5.916	R	0.179	0.087	0.198	0.165	1.278
BU	0.288	0.097	0.571	0.407	4.862						
					f. Pen						
T	0.075	0.752	1.000	0.978	0.330	E	0.120	0.590	1.000	0.966	0.695
MT	0.084	0.729	1.000	0.978	0.372	SI	0.096	0.379	0.944	0.674	1.491
P	0.098	0.686	1.000	0.978	0.458	F	0.033	0.808	0.978	0.978	0.210
K	0.084	0.741	1.000	0.978	0.350	Y	0.109	0.539	0.977	0.933	0.813
C	0.078	0.753	1.000	0.978	0.328	R	0.113	0.298	0.865	0.573	1.903
BU	0.092	0.706	1.000	0.978	0.416						
					g. Achinb						
T	0.222	0.210	0.829	0.723	2.948	E	0.166	0.221	0.723	0.553	2.271
MT	0.216	0.213	0.830	0.681	2.897	SI	0.199	0.128	0.447	0.309	2.492
P	0.213	0.216	0.830	0.660	2.843	F	0.165	0.082	0.287	0.149	2.500
K	0.232	0.205	0.809	0.702	2.946	Y	0.211	0.212	0.734	0.596	2.462
C	0.226	0.210	0.819	0.713	2.900	R	0.120	0.196	0.479	0.351	1.444
BU	0.204	0.219	0.798	0.681	2.644						

[a] The columns list values of the following: D, the disparity eq 6; R_{av}, the average recall eq 3; R_G, the group recall eq 7; R_{GR}, the equivalent rank based group recall; ΔR, the group recall fractional improvement index eq 8. All results were obtained at rank 2000 using the full active set $m = n$.

case (HIVPI) the improvement is substantial. Notably, Wilton et al.,[16] using a related algorithm, also found that ranking by maximum similarity consistently gave the best results. This technique has also been used by Schuffenhauer et al.,[17] who found it to be superior for Unity 2D fingerprints.

DISCUSSION

From its construction it is expected that the group recall should be related in some way to the disparity, D, of the lists that are combined. In particular, the limiting disparity of $1/m$ is reached if all m lists are the same, and the group recall then converges toward the average recall value. More generally, Ng and Kantor[6] have suggested that the effectiveness of data fusion is related to the dissimilarity of two output lists. Since the number of targets chosen varies between groups, correlations with $nD = Z_n(r)/r$, for results that use all available actives ($m = n$), are more meaningful than plots against the disparity itself. This quantity describes the length of the fused list in terms of the length of a single individual list, and it has the range $1 \le nD \le n$. It is proportional to the number of unique compounds that are accessed. Since the fractional improvement for a single list is zero, it is appropriate to consider correlations with $nD - 1$ so that the origin is a constrained point. The first target recovers a number of active compounds proportional to the average recall R_{av}, leaving a number proportional to $1 - R_{av}$

Table 4. a. Average Rank Positions for the Group Recall Results Using Each Activity Class and b. Average Rank Positions Obtained Using the Fractional Improvement Index ΔR For Each Activity Class[a]

similarity coefficient	RTI	Pen	Dop_Ag	Wound-Heal	Hhunity	HIVPI	Achinb	mean
				a.				
T	2.0	4.0	3.5	1.5	1.5	1.0	3.0	2.36
MT	1.0	4.0	3.5	4.0	1.5	3.0	1.5	2.64
P	5.0	4.0	3.5	5.0	4.0	4.0	1.5	3.86
C	3.5	4.0	6.5	1.5	4.0	6.0	4.0	4.21
K	3.5	4.0	6.5	3.0	4.0	5.0	5.0	4.43
BU	6.0	4.0	3.5	6.0	6.0	7.0	6.0	5.50
E	7.0	4.0	1.0	7.0	7.0	8.0	8.0	6.00
Y	8.0	9.0	8.0	9.0	9.0	9.0	7.0	8.43
F	10.0	8.0	11.0	8.0	11.0	2.0	11.0	8.71
SI	9.0	10.0	9.5	10.0	8.0	10.0	10.0	9.50
R	11.0	11.0	9.5	11.0	10.0	11.0	9.0	10.36
				b.				
T	1.0	9.0	4.0	1.0	1.0	4.0	1.0	3.00
MT	2.0	7.0	5.0	3.0	2.0	6.0	4.0	4.14
P	5.0	5.0	6.0	4.0	3.0	1.0	5.0	4.14
K	3.0	8.0	2.0	5.0	4.0	5.0	2.0	4.14
C	4.0	10.0	3.0	2.0	5.0	7.0	3.0	4.86
BU	6.0	6.0	8.0	7.0	6.0	3.0	6.0	6.00
E	7.0	4.0	7.0	6.0	7.0	2.0	10.0	6.14
F	9.0	11.0	1.0	8.0	9.0	8.0	7.0	7.57
Y	8.0	3.0	9.0	9.0	10.0	9.0	9.0	8.14
SI	10.0	2.0	10.0	10.0	8.0	10.0	8.0	8.29
R	11.0	1.0	11.0	11.0	11.0	11.0	11.0	9.57

[a] The results for each coefficient are ranked by the overall mean rank given in the final column.

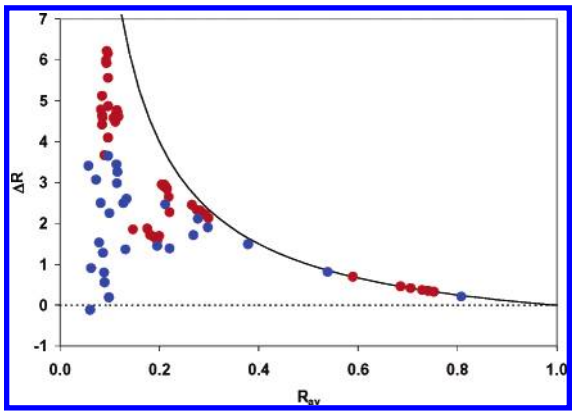

Figure 4. Fractional improvement index ΔR eq 7 plotted against the average recall R_{av}: ● (red) "regular" coefficients; ● (blue) "irregular" coefficients (see text). The full line shows the theoretical maximum that can be obtained.

Table 5. Group Recall Values, R_G, Obtained Using the SUM Fusion Rule Compared with values, R_{GMAX}, Obtained Using the MAX Rule[a]

activity class	μ	R_G	R_{GMAX}	ΔR
Hhunity	0.265	0.520	0.600	5.188
RTIs	0.289	0.681	0.769	6.214
WoundHeal	0.334	0.663	0.783	4.765
Achinbs	0.377	0.829	0.840	2.948
HIVPIs	0.389	0.538	0.817	1.690
Dop_Ag	0.391	0.935	0.976	2.316
Pen	0.525	1.000	1.000	0.330

[a] Final column shows the improvement index obtained from eq 8. Arranged by increasing order of the mean similarity μ.

that may be recovered by fusion with the remaining lists. We might thus expect that fusion of the n lists should lead to an improvement that scales with the product of these

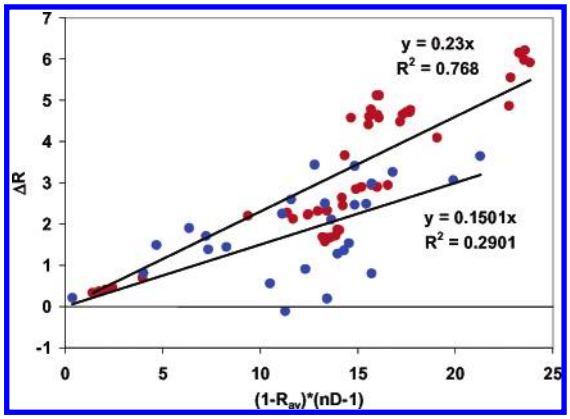

Figure 5. Fractional improvement index ΔR eq 7 plotted against $(1 - R_{av})(nD - 1)$: ● (red) "regular" coefficients; ● (blue) "irregular" coefficients (see text). Linefits shown are constrained through the origin.

quantities:

$$\Delta R \propto (1 - R_{av})(nD - 1) \tag{9}$$

Plots of the measured fractional improvement ΔR against this quantity are presented in Figure 5. The regular and irregular coefficients are again distinguished by color,and independent linefits constrained though the origin are shown.

Results with very low nD are obtained for the penicillins (see Table 3f) and appear in Figure 5 near the origin. This group has a high degree of self-similarity, and thus each target, irrespective of coefficient used, tends to choose similar lists of compounds. For such very self-similar groups values of R_{av} are large (Table 3), and hence values of $(1 - R_{av})$ are small. In addition, because these targets choose many of the same compounds (whether active or not), the disparity of the resulting lists is low and so then is the factor $(nD - 1)$. The converse is true for groups that are more diverse—here R_{av} is small so that $(1 - R_{av})$ is large, but since they tend to choose a wider range of compounds, the disparity is high and $(nD - 1)$ is also high. Thus $(1 - R_{av})$ and $(nD - 1)$ are linked—although this argument does not pretend to show a linear dependence. It is therefore also possible to find passable linear correlations between ΔR and $(nD - 1)^2$ so that a case can also be made for an approximate quadratic scaling of ΔR with the number of unique compounds retrieved. Figure 5 contains the calculated best straight lines for the regular (red) and irregular (blue) coefficients, and it will be seen that the correlation is much stronger for the regular coefficients.

One hypothesis for the relatively low values of R_G obtained using irregular coefficients argues that, whatever target molecule is chosen for a search within a particular group, these coefficients have a predisposition toward choosing similar comparison compounds in each list. The 49 results obtained using regular coefficients have values of nD, which represents the number of unique compounds chosen in the search, with a mean and standard deviation of 18.29 and 5.15, respectively, while the corresponding figures for the 28 results obtained using irregular coefficients are 14.96 and 4.61. These values have a t-statistic of 2.84 which implies that they are significantly different at the 1% level on a two-tailed test. Thus there is some evidence that irregular coefficients yield lower values of nD, indicating that, prior

to merging, the lists contained a higher proportion of the same compounds. This may not be the only origin of the difference between results for regular and irregular coefficients. The results from irregular coefficients in Figure 5 also show a relatively higher degree of scatter, evident from the values of correlation coefficient. Certainly, the distribution of similarity values obtained using the regular and irregular coefficients is different, and this may be another factor contributing to the effectiveness of data fusion.

Subset Size Dependence. We have studied the subset size dependence of group recall (see eq 7) by choosing maximally diverse sets (generated using a Max-Min algorithm with respect to the Tanimoto coefficient) of primary targets from the collections of actives already discussed. The choice of maximally diverse subsets ensures that members of each subset will recover each other with the lowest frequency and that each smaller subset is, for the most part, contained within the larger subsets. Group recall results using the Tanimoto coefficient are shown in Figure 6 plotted against the number of primary targets, m, in the subsets. Values given at $m = 1$ are for the standard average recall.

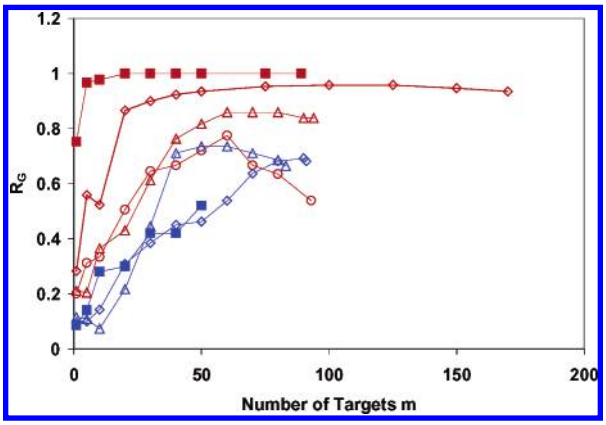

Figure 6. Group recall R_G using subsets of size m vs the number of subset targets m. Groups with high self-similarity (Table 3): ■ (red) penicillin; ◇ (red) Dop_Ag; ○ (red) HIVPI; △ (red) Achinbs; groups with relatively low self-similarity: ◇ (blue) RTI; △ (blue) WoundHeal; ■ (blue) Hhunity. Lines are included only as an aid to the eye.

It will be seen that the group recall generally increases monotonically with m although there are some abrupt reversals at low m. Perhaps using only two or three diverse primary targets generates competition for the fused list that is counterproductive. The group recall plots have the appearance of relaxation curves that reach saturation. In most cases this occurs before all of the targets have been used, and several results actually show a maximum, indicating that the use of too many targets may also be counterproductive. The results are again color coded to distinguish active sets of relatively high and low self-similarity. Those with high self-similarity are seen to have the highest group recall values, but, as would be expected, there is also some evidence that these results reach saturation for smaller subset sizes.

In Figure 7, by analogy with Figure 5, we plot ΔR vs $(1 - R_{av})(mD - 1)$, where mD is the ratio of the total fused length to the length of a single list.

Comparison of this plot with Figure 5 shows that most of the results lie close to the corridor defined by all results for the regular coefficients. However, the correlation is less successful in this case, and there is significant curvature, both

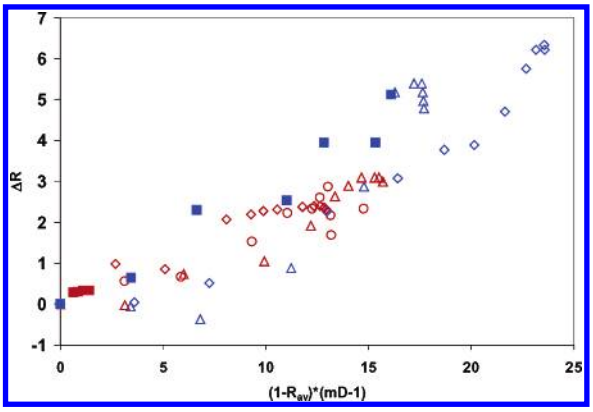

Figure 7. Fractional improvement index ΔR for group recall using subsets of size m plotted against $(1 - R_{av})(mD - 1)$: ■ penicillin; ◇ (red) Dop_Ag; ○ (red) HIVPI; △ (red) Achinbs; groups with relatively low self-similarity: ◇ (blue) RTI; △ (blue) WoundHeal; ■ (blue) Hhunity.

convex and concave, on many of the results for individual activity classes. The correlation appears more successful for the results that have reached saturation.

CONCLUSIONS

This study has shown that there is considerable improvement in the recovery of actives in virtual screening experiments by using a group of structurally related active molecules, rather than a single active molecule. The improvement varies significantly between the various similarity coefficients used and is relatively poor for the set of "irregular" coefficients: the Russell-Rao, Forbes, Simpson, and Yule coefficients. These have been shown previously to have an inherent size bias in the molecules that they judge to be similar to the target structure.[4,10] The degree of improvement scales approximately with the degree of disparity, or difference in composition, between the similarity lists, i.e., the improvement essentially increases with the number of unique compounds in the concatenated list and there is some evidence that this scaling is quadratic. Low values of disparity, resulting from high self-similarity of the activity class or the nature of the coefficient used, lead to relatively poor improvement. A similar correlation of results with the disparity is found when a smaller subset of the activity class is chosen to act as target compounds but the evidence is less compelling. These results also indicate that the group recall values may saturate or reach a maximum before the full activity class size is reached.

From a practical point of view, our studies show that the Tanimoto coefficient is unambiguously the best coefficient to use for the recovery of active compounds using multiple targets, but the reasons for this remain obscure. Using this coefficient, the MAX fusion rule was found to be more effective than the SUM rule for the combination of similarity searches from multiple targets. For practical purposes in virtual screening, we believe that group recall is most advantageous when one has a diverse set of actives. Our results thus suggest that the biggest gain is likely to be obtained from group fusion in the early stages of target recognition, when there may be several candidates with a range of structural types and various modes of action. The

advantages of group recall diminish once a closely related lead series has been identified.

ACKNOWLEDGMENT

We thank Pfizer Global Research and Development for funding; Barnard Chemical Information Ltd., MDL Information Systems Inc., and Tripos Inc. for software and data; and the Royal Society and the Wolfson Foundation for hardware and laboratory support. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council. J.L. thanks Mike Snarey for many discussions regarding the pooling of search fragments.

REFERENCES AND NOTES

(1) Johnson, M.; Maggoria, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.

(2) Buxton, B. F.; Langdon, W. B.; Barrett, S. J. Data fusion by Intelligent Classifier Combination. *Measurement Control* **2001**, *34* (8), 229−234.

(3) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspec. Drug Discov. Des.* **2000**, *20*, 1−16.

(4) 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.

(5) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−440.

(6) Ng, K. B.; Kantor, P. B. Predicting the Effectiveness of Naïve Data Fusion on the Basis of System Characteristics. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 1177−1189.

(7) Wald, L. Some Terms of Reference in Data Fusion. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1190−1193.

(8) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(9) The BCI software is available from Barnard Chemical Information Ltd. at URL http://www.bci.gb.com/.

(10) 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.

(11) Wang, R.; Wang, S. How Does Consensus Scoring Work For Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(12) Mazzatorta, P.; Benfenati, E.; Neagu, D.; Gini, G. The Importance of Scaling in Data Mining for Toxicity Prediction. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1250−1255.

(13) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343−357.

(14) The MDL Drug Data Report database is available from MDL Information Systems at URL http://www.mdli.com/.

(15) The *Dictionary of Natural Products* database is available from Chapman & Hall/CRC at URL http://www.crcpress.com/.

(16) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.

(17) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

CI049867X