Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion[†]

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

Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, U.K.

Jens Loesel

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

Received December 20, 2004

In a recent companion paper we have related the operation of simple data fusion rules used in virtual screening to a multiple integral formalism. In this paper we extend these ideas to the analysis of data fusion methods applied to real data. We examine several cases of similarity fusion using different coefficients and different representations and consider the reasons for positive or negative results in terms of the similarity distributions. Results are obtained using the SUM-, MAX- MIN-, and CombMNZ-fusion rules. We also develop a customized fusion rule, which provides an estimate of the optimal possible result for fusing multiple searches of a specific database; this shows that similarity fusion can, in principle, achieve retrieval enhancements even if this is not achieved in practice with current fusion rules. The methods are extended to analyze the comparatively successful results of group fusion with multiple actives, and we provide a rationale for the observed superiority of the MAX-rule over the SUM-rule in this context.

INTRODUCTION

Data fusion is being increasingly used to enhance the performance of systems for virtual screening in pharmaceutical and agrochemical research.[1,2] Of the available virtual screening techniques, similarity searching is the simplest and computationally cheapest to perform. The method takes one or more known bioactive molecules, often known as the *target* or *reference structure*, and searches a database to find those structures that are most similar; these are the most likely to exhibit similar properties to those of the reference structure and hence are candidates for acquisition and biological testing. Different similarity measures can yield different lists of nearest neighbor compounds considered similar to the reference structure.[3,4] Figure 1 sketches the lists that might be recovered from two searches using the same reference structure and different similarity measures *x* and *y*. The recovered compounds that are members of the same activity class as the reference structure are shown as red bars and the gray regions represent other recovered but nonactive compounds. A database may typically contain 10^5-10^6 compounds, but perhaps only $\sim10^3$ of the most similar compounds are required for expensive physical screening, and so the lists are truncated at the required rank.

The known active compounds that are recovered are called true positives (*TP*), and the recovered nonactive compounds are called false positives (*FP*). Actives remaining to the right of the truncation point are called false negatives (*FN*), and the remaining nonactives are the true negatives (*TN*). The effectiveness of the retrieval technique can then be assessed by computing the *cumulative recall*, *R*, which is the fraction of known active compounds that is recovered at the truncation rank[5] *N*. The *precision P* is an alternative measure defined as the fraction of actives recovered relative to all compounds recovered, i.e., the truncation rank.

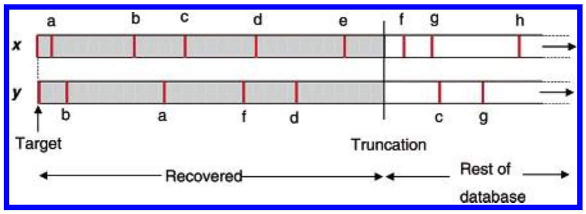

Figure 1. Sketch of retrieved lists for a pair of similarity measures *x* and *y*. The compounds most similar to the reference are at high rank to the left. Labeled red bars represent individual members of a single activity class, and the gray regions represent other compounds recovered from the database above the truncation rank. Compounds to the right of the truncation point are of lower similarity and not recovered in the final lists.

$$R(N) = TP/(TP + FN) \tag{1}$$

$$P(N) = TP/(TP + FP) \tag{2}$$

By using each member of the activity class as a reference compound one can determine an average recall value. This average may vary considerably between different activity classes, and several determinations are needed to assess the overall effectiveness or systematic dependencies of a given technique.

As illustrated in Figure 1 the rank order of the compounds retrieved by the two measures for a given reference may be different, and indeed some of the compounds (whether active or nonactive) recovered by one measure may not be retrieved by the other at a given truncation rank. We call these *unmatched* compounds or similarity values; examples are compounds *c*, *e*, and *f* in Figure 1. Conversely, compounds such as *a*, *b*, and *d* are termed *matched* since they appear in both truncated lists. By using data fusion techniques to

[†] Dedicated to Professor Johann Gasteiger.

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

10.1021/ci0496144 CCC: $33.50 © 2006 American Chemical Society
Published on Web 08/29/2006

Table 1. MDDR Compound Classes Used in This Study[a]

activity class	abbreviation	class	*n*	*m*	*s*
reverse transcriptase inhibitors	RTI	agents for AIDS	91	0.289	0.139
wound healing agents	WDH	dermatological agents	83	0.334	0.189
acetyl-cholinesterase inhibitors	ACH	agents for cognition disorders	94	0.377	0.144
HIV1 protease inhibitors	HIV	agents for AIDS	93	0.389	0.155
dopamine autoreceptor agonists	DAG	antipsychotics	170	0.391	0.172
penicillins	PEN	β-lactam antibiotics	89	0.525	0.171

[a] Each contained a random selection of *n* reference structures from the complete activity classes. The final columns show the mean μ and standard deviation σ for pairwise similarities for the group obtained using the Tanimoto coefficient and BCI fingerprints.

combine the nearest neighbor lists obtained using two or more different methods, the goal is to obtain a single new list that returns higher values of recall and precision than the original lists at the same rank, i.e., an increased level of enrichment in the actives.

Studies of data fusion applied to similarity searching fall into three types. The combination of the nearest neighbor lists obtained using different similarity measures is called *similarity fusion*, and this can be further subdivided into the use of the same similarity coefficient with different representations (e.g., ref 2) and the use of the same representation with different similarity coefficients (e.g., ref 6). The objective of these studies has been to identify combinations of measures that reliably lead to retrieval enhancement, but the results of such experiments have been inconsistent. More recently, we have applied data fusion methods to the combination of similarity lists obtained from multiple reference structures. In this method, called *group fusion*, only a single measure of similarity is used, and the resulting lists that are generated by each of the different reference structures are combined. In this case the lists *x* and *y* in Figure 1 would represent results obtained using the same similarity measure with different reference structures belonging to the same activity class. In marked contrast to similarity fusion, group fusion has led to reliable and consistent results for a range of activity classes.[7,8] The main objective of the study reported here is to probe the reasons for this marked difference in behavior with a view to finding ways of improving the results of similarity fusion.

In a companion paper, we have examined the mechanism of data fusion when applied to the combination of similarity values using simple model distributions.[9] There, we have shown how simple data fusion rules can be represented by multiple integration of the similarity frequency distributions obtained by different search methods. Using model systems we found that the success or otherwise of data fusion depends on differences between the recovered-active (i.e., *TP*) and recovered-nonactive (i.e., *FP*) similarity distributions and how they relate to the appropriate integration regions. In the present paper we draw on these theoretical results to examine how data fusion interacts with real data and what this can tell us about the reasons for its heretofore enigmatic behavior. We examine the three possible types of fusion in turn: (i) same reference structure, same representation, different similarity coefficient; (ii) same reference structure, same similarity coefficient, different representation; and (iii) same representation, same similarity coefficient, different reference structures (group fusion). Several different fusion rules are described, and we also outline a customized fusion rule that provides an estimate of the maximum performance possible from a data-fusion system.

Table 2. Similarity Coefficients Used in This Work[a]

coefficient	abbrev	expression
Tanimoto	T	$S_T = c/(a + b - c)$
modified Tanimoto[13]	MT	$S_{MT} = S_T(2 - \rho_0)/3 + S_{T0}(1 + \rho_0)/3$
Cosine	C	$S_C = c/\sqrt{ab}$
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	$S_P = (Nc - ab)/$ $\sqrt{Nab(N - b)(N - a)}$
Squared Euclidean	E	$S_E = (a + b - 2c)/N$
Modified Russell-Rao[14]	R	$S_{MR} = c/a$
Modified Forbes[14]	F	$S_{MF} = c/b$
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 reference-structure 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 density of set bits ρ_0 for the fingerprints and $S_{T0} = d/(N - c)$ for the Tanimoto coefficient for absent features.

METHODS

Measurement of Search Performance. The pharmacologically active compound classes used in this work were all selected from the *MDL Drug* Data Report (MDDR)[10] and are collected in Table 1. The version of the database used here contained 102 443 compounds, and the activity classes used form a subset of those studied by Whittle et al.[8] The molecular structures were represented by fragment bit-strings generated using BCI[11] and Pipeline Pilot[12] software. Most of the results presented use WDH activity class, but comparable results have been obtained with the other activity classes listed in Table 1. Structures are compared using one of several similarity coefficients, $S(A,B)$ (as listed in Table 2), that quantify the degree of resemblance between the reference structure *A* and a *comparison structure B* in the database that is being searched (i.e., MDDR in this case).

The results of a similarity search are first linearly rescaled to map the original values onto the range 0−1. Thus, for each unscaled similarity result $S(A,B)$ a rescaled value $S^*(A,B)$ is obtained using

$$S^*(A,B) = \frac{S(A,B) - S_{\min}(A)}{S_{\max}(A) - S_{\min}(A)} \tag{4}$$

where $S_{\max}(A)$ and $S_{\min}(A)$ are the maximum and minimum values in the ranked list for reference *A*. In similarity fusion, of type (i) or (ii) discussed in the Introduction, the values obtained using two or more measures are combined using one of several fusion rules. For *m* lists of *N* scaled similarity values $S_k^*(i,j)$ relating targets i and comparison structures j using similarity measures k, we compute a fused score $S_{FUS}(i,j)$ for each recovered structure from

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

If structure j is not found in one of the truncated lists the scaled similarity is assumed to be zero. The fusion rule, F, indicates the method used to combine the similarities. Whittle et al. have analyzed four such rules:[9] for the SUM-rule a summation sign, Σ, replaces F; for the MAX-rule F operates on the set of similarities to choose the maximum value for each j; for the MIN-rule F operates to choose the minimum value for each j; and the CombMNZ-rule additionally includes a weighting term such that the SUM-rule score for each retrieved structure is multiplied by the number of times that structure appears in any of the truncated lists.

In all cases the structures j are then ranked according to the returned values of the combined similarity. However, because different measures may retrieve partially different comparison compounds the fused list will usually be longer than any of those obtained by the individual measures. Thus, to make a fair comparison the results must all be compared at the same rank. In a practical sense, this corresponds to the choice of a limited number of compounds, enriched by virtual screening, to present for real screening experiments. The *enhancement*, ϵ, obtained by data fusion can be expressed as the fraction

$$\epsilon = \frac{R_z - \max\,(R_x,R_y)}{\max\,(R_x,R_y)} \tag{6}$$

where R_x and R_y are the recall values obtained using the single-measures x and y, and R_z is the fused recall value.

Group fusion, case (iii) discussed in the Introduction, applies analogous methods to the lists obtained using n related reference structures and a single measure.[8] For each of the comparison structures j that are found in the ranked lists we compute the fused result $S_{GF}(j)$ from the fusion rule F and the scaled similarities, $S^*(i,j)$ for a particular coefficient, as obtained from eq 4

$$S_{GF}(j) = F_{i=1}^{n}[S^*(i,j)] \tag{7}$$

In this case the operation is carried out over all n lists from each of the reference compounds i so that $S_{GF}(j)$ represents a combined similarity measure of compound j to all reference molecules used to generate the lists.

Precision Maps. The scaled similarity values associated with any of the lists obtained in a similarity search can be used to obtain frequency distributions, effectively proportional to the occurrence probability of a given similarity value. Moreover, these can be split into components arising from recovered-active (TP) and recovered-nonactive (FP) contributions. For single measures, some form of bias toward high similarity in the recovered-active distribution relative to the recovered-nonactive distribution lies at the root of enrichment, i.e., recall values, above the random expectation. When two such lists are combined, the results can be correspondingly represented in two-dimensions, as shown in Figure 2. From this, a pair of recovered-active and recovered-nonactive bivariate similarity distributions can be defined.[9]

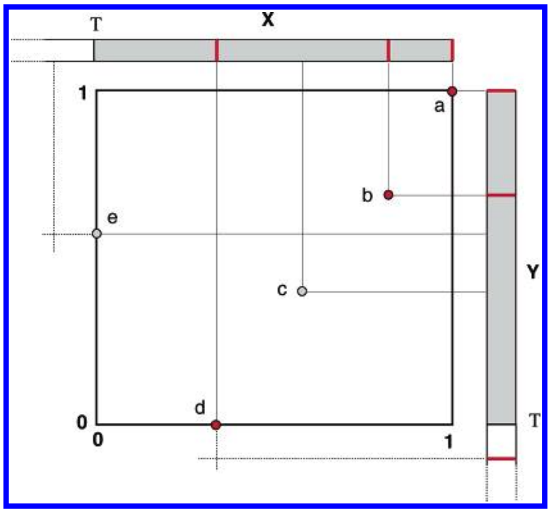

Figure 2. Schematic representation of the combination of two scaled similarity lists X and Y to form a two-dimensional field. As in Figure 1, the similarity lists are shown as bars with red bars marking the positions of recovered-actives, gray regions marking recovered-nonactives. Compounds with similarities below the truncation point, T, are unrecovered and assigned a similarity of zero. The schematic shows the following: (a) the target; (b) an active recovered by both measures; (c) a nonactive recovered by both measures; (d) an active recovered by X but not Y; and (e) a nonactive recovered by Y but not X.

The comparison of any pair of truncated nearest-neighbor lists obtained using different measures will in most cases include some values that relate to structures unique to just one of the lists. Examples of these unmatched values are shown as points d and e in Figure 2, where they reside on the axes because missing values are assigned a value of zero. We define the *match ratio*, M_R, for pairs of lists as

$$M_R = \frac{\text{no. of structures common to both lists}}{\text{total no. of different structures in both lists}} \tag{8}$$

The value of this Tanimoto-like coefficient, computed on a pair of lists, varies between 0 (for lists that have no structures in common) and 1 (for lists that contain exactly the same compounds). We will usually quote this as an average, $\bar{M}_R$, taken over all pairs of lists considered.

By superimposing the results from all reference structures used, an average bivariate distribution appears as a variable density of points. In many cases there is a discernible difference between the recovered-active and recovered-nonactive distributions, and it is this which can be exploited by data fusion. An example is shown in Figure 3(a), where the points represent scaled similarity values obtained using the scaled Forbes coefficient plotted against those obtained using the scaled Russell-Rao coefficient for the set of 83 wound healing agents. In this case, the recovered-active points (red) are more evenly spread than the recovered-nonactive points (gray), which are concentrated toward lower similarity for both measures. Thus the recovered-active distribution is biased toward high similarity for both measures relative to the recovered-nonactive distribution. Because the recovered-nonactive pairs significantly outnumber the recovered-active pairs, they were randomly sampled for display and only about 1 in 10 are shown. The number of points in each of the grid squares shown is then proportional to the

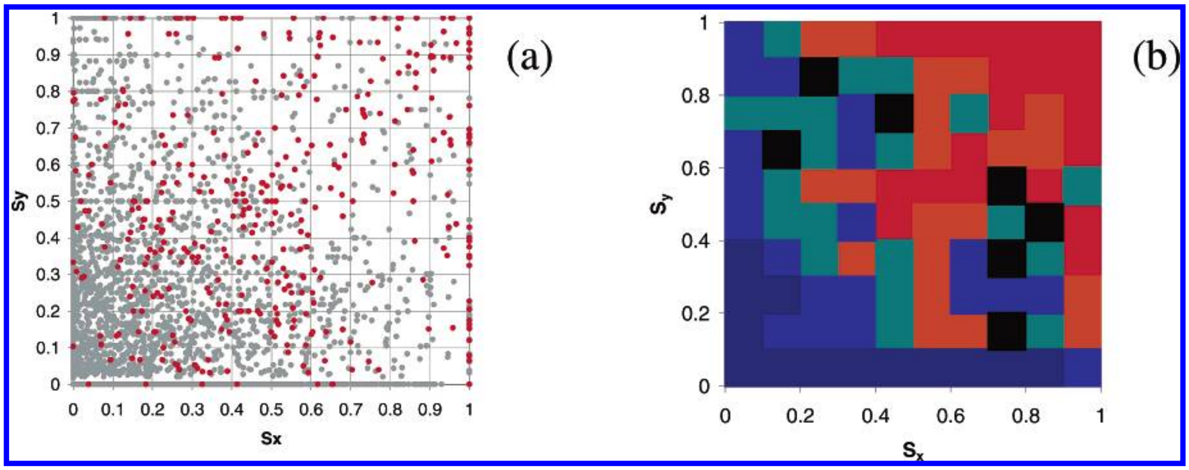

Figure 3. (a) Paired scaled similarity values S^*_{ij} for the scaled Forbes (S_x) and Russell-Rao (S_y) coefficients obtained using 83 wound healing agents from the MDDR as the active set. Results were taken down to rank 2000, and the whole of the MDDR (less the 83 known agents) was used as the nonactive set. (red circle) recovered-active values; (black circle) recovered-nonactive values (only 10% of these points are shown for clarity). (b) Corresponding map showing the local precision or active fraction. Color-coded local precision ranges: (red square) $P > 0.8$; (orange square) $0.8 \geq P > 0.6$; (green square) $0.6 \geq P > 0.4$; (blue square) $0.4 \geq P > 0.2$; (black square) $0.2 \geq P > 0.0$; (black square) $P = 0$; (gray square) *neutral* − no compounds recovered.

average density over each square: the plot is thus essentially a graphical representation of the *partial* bivariate densities $\phi(x,y)$ for the n recovered-active pairs in red and $\Phi(x,y)$ for the $(N-n)$ recovered-nonactive pairs in gray, where x and y are similarity scales and N compounds are recovered. These are defined[9] so that $n = N\int_0^1 \mathrm{d}y \int_0^1 \phi(x,y)\mathrm{d}x$ and $(N-n) = N \int_0^1 \mathrm{d}y \int_0^1 \Phi(x,y)\mathrm{d}x$. We also define $\Phi_T(x,y) = \Phi(x,y) + \phi(x,y)$ as the bivariate density for all recovered compounds, where $\int_0^1 \mathrm{d}y \int_0^1 \Phi_T(x,y)\mathrm{d}x = 1$.

In our companion paper[9] we showed that the fused recall for such bivariate distributions could be represented by integration of the joint similarity distribution over a region appropriate to the fusion rule. It can be seen from Figure 3(a) that the ratio of recovered-actives to recovered-nonactives varies with position, and so the region of integration used can have a significant influence on the precision P; the precision is defined in this case as the fraction of active compounds, $n(\mathrm{Q})$, to total compounds, $N(\mathrm{Q})$, retrieved over a given region Q. This can be expressed as

$$P(\mathrm{Q}) = \frac{n(\mathrm{Q})}{N(\mathrm{Q})} = \frac{\int\int_{\mathrm{Q}} \phi(x,y)\mathrm{d}x\mathrm{d}y}{\int\int_{\mathrm{Q}} \Phi_T(x,y)\mathrm{d}x\mathrm{d}y} \qquad (9)$$

where x and y are similarity scales. $P(\mathrm{Q})$ is then a measure of how efficient a given region is at accumulating actives. The denominator is the cost of using the region in terms of rank, while the numerator gives the return in terms of actives retrieved.

To compare the results of data fusion with that of a single measure we need to find the number of active compounds recovered by each measure at the same rank and thus obtain the enhancement ϵ as

$$\epsilon = \frac{n(\mathrm{Q}) - n(\mathrm{S})}{n(\mathrm{S})} \text{ for Q,S:} N(\mathrm{Q}) = N(\mathrm{S}) \qquad (10)$$

where S is the single measure integration region. Since the latter condition is specified we can divide through each term by either $N(\mathrm{Q})$ or $N(\mathrm{S})$ to obtain explicitly

$$\epsilon = \frac{P(\mathrm{Q}) - P(\mathrm{S})}{P(\mathrm{S})} \text{ for Q,S:} N(\mathrm{Q}) = N(\mathrm{S}) \qquad (11)$$

Fusion enhancement hence depends explicitly on differences in the precision between the regions used to recover compounds using data fusion and the regions used to recover compounds using a single measure. This means that the best region of integration to use is that which best covers the areas containing values associated with actives while avoiding those that contain nonactive values. One way of trying to ensure this is to cover the area with a grid, as shown in Figure 3a, and to score each grid square according to the local precision or active fraction

$$P(g) = n(g)/N(g) \qquad (12)$$

where $n(g)$ is the number of active values in grid square g, and $N(g)$ is the total number of values in that square—effectively the cost in terms of rank of including it in a summation. The squares are then ranked with the highest score top and split into a number of groups according to value. By coloring each of these categories appropriately we obtain a *precision map* that identifies those regions having the richest seams of active compounds relative to the cost of mining in terms of rank. This is done in Figure 3b for the same data as Figure 3a. We have used five categories for nonzero precision: $P(g) > 0.8$; $0.8 \geq P(g) > 0.6$; $0.6 \geq P(g) > 0.4$; $0.4 \geq P(g) > 0.2$; $0.2 \geq P(g) > 0.0$. Areas corresponding to zero precision, $P(g) = 0.0$, are rendered black; these contain only nonactive values and are best avoided by any fusion scheme. Any areas colored gray in the figures contain no values at all; they are therefore neutral regions for any fusion rule and formally have an undefined precision value. The results shown in Figure 3(b) and other precision maps presented here were obtained without the sampling of nonactive values used to display Figure 3(a) and thus represent the "true" relative precision values. In the remainder of the paper we address the question of how best to harvest the high-precision regions so as to maximize the recovery of actives at a given rank.

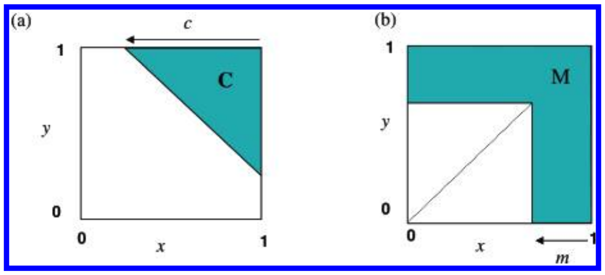

Figure 4. Sketch of two-dimensional regions appropriate for the fusion of two similarity lists represented by the values *x* and *y*. For (a) the SUM-rule, C corresponds to the collection region, and the variable *c* represents the combined similarity. For (b) the MAX-rule, M corresponds to the collection region, and the variable *m* represents the combined similarity.

ANALYSIS OF SIMILARITY FUSION

Similarity Fusion Using Different Coefficients. In this section we examine similarity fusion results using the model developed previously,[9] where we showed that integration regions over the joint similarity distributions could represent simple fusion rules. The regions used by the SUM and MAX rules to collect values from a bivariate distribution are represented diagrammatically in Figure 4, where the unit square contains points representing the similarities of all compounds available in the fused list. As the regions extend to fill the unit square, more compounds are collected at lower combined similarity and the rank increases accordingly.

We performed a number of similarity fusion experiments using the activity classes given in Table 1 and the coefficients given in Table 2. For each activity class, results were combined from the best performing individual coefficients and also some pairs that have given positive results in previous studies that we have carried out.[6,15] Similarity values were scaled using eq 4, and the results were combined using the SUM and MAX rules according to eq 5. Out of 70 results obtained we found that data fusion by this method was generally unsuccessful, giving only three clearly positive results when compared with the best overall single measure. We are interested in probing the reasons for this success and therefore focus attention on the WDH activity class that gave the best positive fusion result. This was obtained using a combination of results obtained with the Forbes and Squared Euclidean coefficients. However, from the standpoint of this analysis it will be convenient to examine other combinations first.

Successful data fusion is frequently said to arise from the combination of results from equally good but different measurements. The Forbes and Russell-Rao coefficients provide very different expressions of similarity, and they might therefore be good candidates for fusion.[14] When data fusion was applied to the data set used to generate Figure 3 (WDH with the Forbes and Russell-Rao coefficients), the SUM-rule improved on the retrieval obtained by using the best coefficient alone (the Forbes) by 5.96% at rank 1000, and the MAX-rule gave an improvement of 2.58%. These are effectively values of the enhancement, eq 6, expressed as a percentage. Although these results were themselves surpassed by using other single coefficients, including the Tanimoto, taken in isolation this example does illustrate a case of successful data fusion and the associated bivariate distribution of similarity values that gave rise to it. This result would appear to support our hypothesis since the areas of high precision in Figure 3 correspond quite well with the upper triangular shape of the SUM-rule integration area.

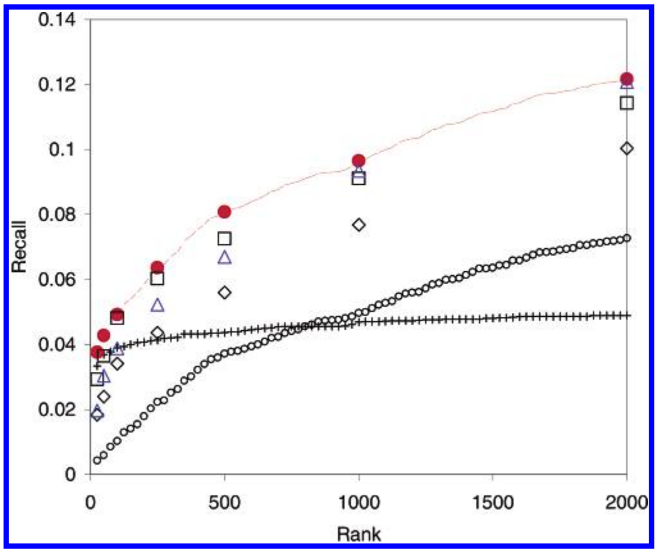

Figure 5. The average cumulative recall obtained using 83 wound healing agents from the MDDR as the active set. For selected values of rank: ◇ Russell-Rao; □ Forbes; (red circle) fusion using the SUM-rule: △ fusion using the MAX-rule. At rank intervals of 25: + SUM-rule fused recall from matched entries; ○ SUM-rule fused recall from unmatched entries; (red dotted line) total fused recall using the SUM-rule.

The relative importance of the matched and unmatched contribution in data fusion has been the subject of some debate recently,[16] and we have also examined these contributions in our companion paper.[9] Within a fusion routine it is a straightforward matter to tag those values in the fused list that are generated from matched entries in the original lists and separate them from those that are generated from unmatched entries. It is thus possible to separate the recall into portions arising from the matched entries, $R^M(r)$, and from the unmatched entries $R^U(r)$.

$$R(r) = R^M(r) + R^U(r) \tag{13}$$

These are shown in Figure 5 (for the data represented in Figure 3) where it can be seen that the matched recall is the major contributor at the top of the ranking but that the unmatched recall makes up the major proportion of the total recall at rank 1000 and beyond. The average match ratio for all recovered values (actives and nonactives) is 0.069 at rank 1000 (see Table 3), i.e., only ca. 7% of all recovered values (active and nonactive) are matched. Most entries must therefore lie on one of the axes of Figure 3(a) because the scaled similarity value is zero for one coefficient in an unmatched pair. Thus in terms of the SUM-rule integration region, as shown in Figure 4a, the diagonal is reached at a very early stage because the body of the diagram represents relatively few entries. However, just over half ($\bar{M}_R = 0.513$, Table 3) of the available active compounds are matched at rank 1000 on average, which explains the relatively steep rise (for rank < 100) of the matched component of the recall in Figure 5. Note that the ratio of recall values at this point differs from $\bar{M}_R$ because it is a ratio of averages and not an average of ratios.

As a general rule, it is thought that dissimilar output and comparable performance is a recipe for potentially effective fusion. In an effort to quantify this heuristic, Ng and Kantor

Table 3. Values of Product-Moment Correlation Coefficient between the SUM- and MAX-Rule Lists and Fusion Enhancement for Some Coefficient Combinations (See Table 2 for Coefficient Abbreviations) Using the Wound Healing Agent Activity Class[a]

	all values				recovered-active				enhancement	
combination	$\bar{M}_R$	r	ρ_u	ρ_m	$\bar{M}_R$	r	ρ_u	ρ_m	ϵ(SUM)	ϵ(MAX)
T_BU	0.807	0.964	−0.444	0.964	0.937	0.994	−0.287	0.994	−0.0031	−0.0094
T_C	0.902	0.987	−0.455	0.987	0.960	0.997	−0.581	0.997	−0.0047	0.0031
T_E	0.535	0.826	−0.456	0.858	0.814	0.989	−0.292	0.991	−0.0047	0.0031
T_F	0.293	0.267	−0.428	0.440	0.69	0.857	−0.198	0.843	0.0785	0.0785
T_K	0.785	0.954	−0.444	0.957	0.909	0.994	−0.571	0.995	0.0000	0.0000
T_MT	0.926	0.997	−0.456	0.997	0.961	1.000	−0.333	1.000	0.0016	0.0016
T_P	0.838	0.966	−0.440	0.966	0.933	0.995	−0.283	0.995	0.0016	0.0016
T_R	0.279	0.250	−0.397	0.473	0.665	0.812	−0.478	0.847	−0.0392	−0.0487
T_SI	0.253	0.225	−0.422	0.411	0.710	0.799	−0.438	0.832	−0.0031	−0.0204
T_Y	0.565	0.641	−0.461	0.731	0.834	0.819	−0.460	0.861	0.0030	0.0000
C_BU	0.780	0.956	−0.460	0.958	0.939	0.998	−0.331	0.998	0.0110	0.0094
E_BU	0.649	0.897	−0.484	0.908	0.855	0.994	−0.137	0.996	−0.0016	0.0491
E_C	0.534	0.816	−0.475	0.853	0.83	0.989	−0.188	0.992	0.0079	0.0189
E_F	0.166	0.063	−0.463	0.418	0.626	0.813	−0.253	0.799	0.1359	0.1456
E_R	0.531	0.479	−0.409	0.553	0.794	0.840	−0.368	0.843	−0.0210	−0.0307
E_SI	0.247	0.214	−0.466	0.391	0.722	0.796	−0.387	0.793	0.0405	0.0453
E_T	0.535	0.826	−0.465	0.858	0.814	0.989	−0.292	0.991	−0.0047	0.0031
F_Y	0.246	0.144	−0.483	0.414	0.672	0.749	−0.292	0.737	0.0469	0.0499
R_F	0.069	−0.247	−0.440	0.290	0.513	0.599	−0.367	0.606	0.0597	0.0258
R_SI	0.411	0.511	−0.484	0.709	0.704	0.834	−0.361	0.855	0.0470	0.0643
Y_SI	0.463	0.617	−0.473	0.649	0.829	0.882	−0.129	0.878	−0.0242	−0.0030
Y_K	0.638	0.728	−0.451	0.781	0.891	0.857	−0.531	0.868	0.0030	0.0000

[a] Average values $\bar{M}_R$ of the match ratio (eq 8) and total Product-Moment Correlation values, ρ (eq 14), were obtained for lists indicated in the top row. Averages for recovered-actives are taken over nonempty lists only. The values ρ_u and ρ_m were similarly obtained using only unmatched and matched values from the same lists, respectively. Results are compared with the fusion enhancements relative to the best individual recall obtained for each pair at rank 1000 using SUM and MAX rules given in the final columns.

have suggested that Kendall's τ may be used as an indicator of effective fusion; this statistic is a measure of rank correlation with small or negative values indicating dissimilar output.[17] For continuous values, the Product-Moment Correlation Coefficient, ρ (also known as Pearson's Correlation Coefficient), performs an analogous role and is also considerably faster to compute

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad (14)$$

where x_i and y_i are similarity values with distribution means $\bar{x}$ and $\bar{y}$. Accordingly, for the WDH data we collect values of this statistic in Table 3 for a range of coefficient combinations and compare with values of the enhancement eq 6 obtained using the SUM and MAX rules. We have ascertained[9] that data fusion is driven by differences between recovered-active and recovered-nonactive distributions. The latter is almost the same as that for the complete list (which might be more generally accessible) for small active fractions. We therefore give values for whole lists (where the i in eq 14 includes all molecules in the list) and also for the recovered-active contribution (where the i in eq 14 refers only to the active comparison molecules). For low active fractions the recovered-nonactive component is very close to that obtained for evaluations using the complete lists. These components are each further split into contributions from the matched and unmatched values. Average values of the match ratio are also included. Note that some list combinations were empty of recovered-actives and hence have undefined values of ρ and the match ratio. For these, averages were taken over the nonempty lists only. As the unmatched values reside on the axes of a bivariate plot they reside mostly in the second and fourth quadrants relative to the means and thus make a negative contribution to ρ. In this case the contribution is roughly constant. Values of ρ for the matched values are all positive for these results and increase with match ratio. Values of the overall ρ are intermediate at low match ratio, but, as expected, they become equivalent to the matched result as the match ratio increases. Thus although the negative value of the overall ρ for the Russell-Rao/Forbes combination might suggest that the results are negatively correlated, this is seen to be largely due to the effect of unmatched values.

The relatively even distribution of similarity values in Figure 3a for the Russell-Rao and Forbes combination is reflected in the very low value of 0.29 obtained for ρ_m. In contrast, the results from many of the combinations are strongly correlated and return relatively high values of ρ_m. For example, strong correlations are found for the Tanimoto-Cosine and the Tanimoto-Baroni-Urbani combinations. Lower values are generally returned by any combination that includes the Russell-Rao, Forbes, Simpson, and Yule; these are coefficients that we have called "irregular" since they are well-known to behave differently from the others.[14] The Squared Euclidean coefficient occupies a halfway house and shows some of the characteristic behavior of this group; it has a tendency to choose larger structures, for example. However, the Squared Euclidean coefficient also frequently gives good recall results as a single measure and when paired with a regular coefficient, such as the Tanimoto, gives significantly different lists, although not as different as those obtained using an irregular coefficient. The overall match ratio and ρ values for the Tanimoto-Squared Euclidean combination are thus relatively low when compared with the

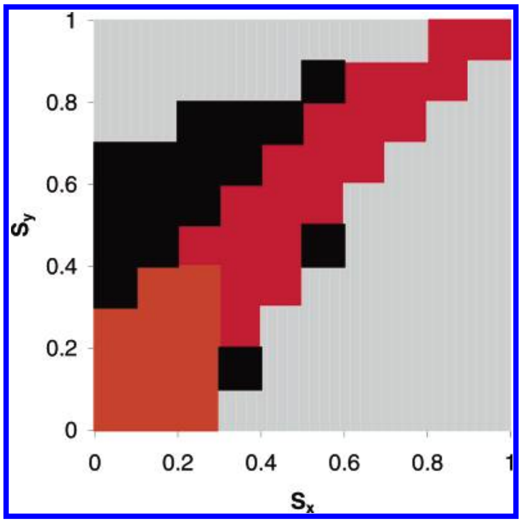

Figure 6. Precision map for the scaled Tanimoto and Squared Euclidean coefficients obtained using 83 wound healing agents from the MDDR as the active set. Results were taken up to rank 2000, and the whole of the MDDR (less the 83 known agents) was used as the nonactive set. Results. Color-coded precision ranges: (red square) $P > 0.8$; (orange square) $0.8 \geq P > 0.6$; (green square) $0.6 \geq P > 0.4$; (blue square) $0.4 \geq P > 0.2$; (black square) $0.2 \geq P > 0.0$; (black square) $P = 0$; (gray square) *neutral* − no compounds recovered.

Tanimoto-Cosine pairing for example. Thus, a pairing of measurements using the Tanimoto and Squared Euclidean coefficients would seem to have a good chance of success. For the WDH data the difference between the fused and single-measure results is less than ±1% for the SUM and MAX rules, and the precision map is shown in Figure 6. Despite the lower statistical values obtained for ρ_m, the data are strongly grouped along the diagonal.

Significantly, it can now be seen that neither of the fusion-rule shapes shown in Figure 4 matches the shape of the distribution shown in Figure 6. Specifically, neither region matches the zone of high precision retrieval any better than would a rectangular region parallel to one of the axes, corresponding to single measure recall.[9] This suggests that neither fusion rule will be particularly effective compared with the single measures. Indeed we find in practice that there is no significant enhancement obtained for this activity class. Some of the other activity classes produced small (∼2%) improvements for this combination at rank 1000, but these were rank-dependent and not reproducible across all classes.

We have previously shown analytically[9] that a difference in the degree of correlation between the recovered-active values and recovered-nonactive values for two lists of similarity values does influence the results of data fusion. However, positive or negative enhancement may be obtained depending on rank, and, for a given difference, the effect is reduced if the degree of correlation is high for both sets of data. For our experimental results, the enhancement obtained for all of the coefficient combinations of Table 3 is plotted against differences in correlation obtained for the matched, $\Delta\rho_m$, and unmatched contributions, $\Delta\rho_u$, in Figure 7. There is no direct dependence of enhancement on either factor, although there is an increase in the range of enhancement values as the correlation difference increases. This is not inconsistent with our theoretical results, since they suggest that either positive or negative enhancement values could be obtained depending on the rank. Although our results are taken at fixed rank, the position is complicated because the position of changeover between positive and negative enhancement is also affected by the match ratio. In addition, other factors, such as similarity bias, may be contributing to the results.

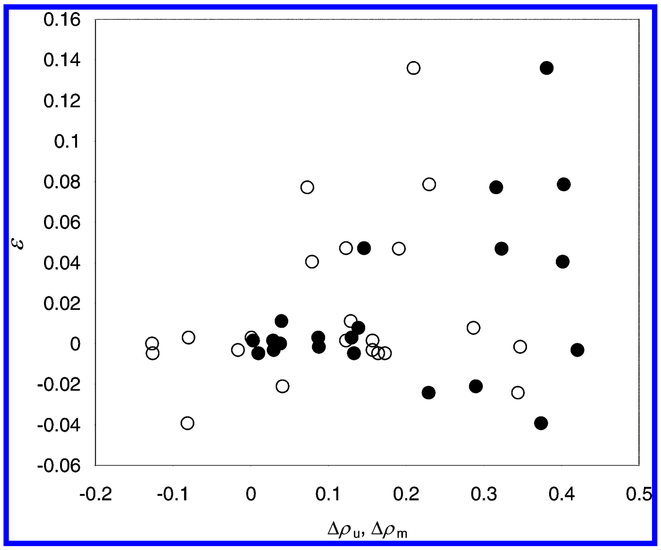

Figure 7. The enhancement, ϵ, obtained using the SUM-rule for the coefficient combinations of Table 3 plotted against the difference in the Product-Moment Correlation Coefficient between recovered-active and recovered-nonactive values: $\Delta\rho_u$ ○ unmatched and $\Delta\rho_m$ ● matched.

Since one combination in particular, the Squared-Euclidean and Forbes coefficients, gave significant improvement over the best single value recall obtained by any of the coefficients, it is worth examining this pairing in more detail. The precision map for this pairing is illustrated in Figure 8, which shows high levels of precision over a broad range but the shape is not so clearly triangular as that seen in Figure 3b. In fact for this pairing the MAX-rule gave a slightly better result than the SUM-rule (Table 3). A significant block of high precision pixels appears across the top of this diagram that would be coincident with one arm of the MAX-rule region (Figure 4b) and also benefits recovery using the Forbes coefficient. Along the right-hand edge there is a significant region of zero or "negative" precision, which would benefit neither the MAX-rule nor retrieval by the single measure, in this case the Squared Euclidean coefficient. However, the main arm of the high-precision results is roughly aligned with this edge, and it would seem that this taken together with the high precision region along the top of the diagram is enough to give the MAX-rule an advantage in this case.

It would clearly not be possible from the data here to predict successful fusion in this case, and the arguments we have given cannot be claimed robust. Indeed, the precision map for the Tanimoto and Russell-Rao coefficients is only subtly different to that for successful pairing of Squared Euclidean and Forbes, but the enhancement for both SUM and MAX-rules is negative at rank 1000. We believe that such differences are due to the fact, noted previously, that combination by data fusion involves so many factors; the fusion of just two sets of similarities depending upon eight distinct but interdependent distributions: the recovered-active similarities obtained using the two measures include matched

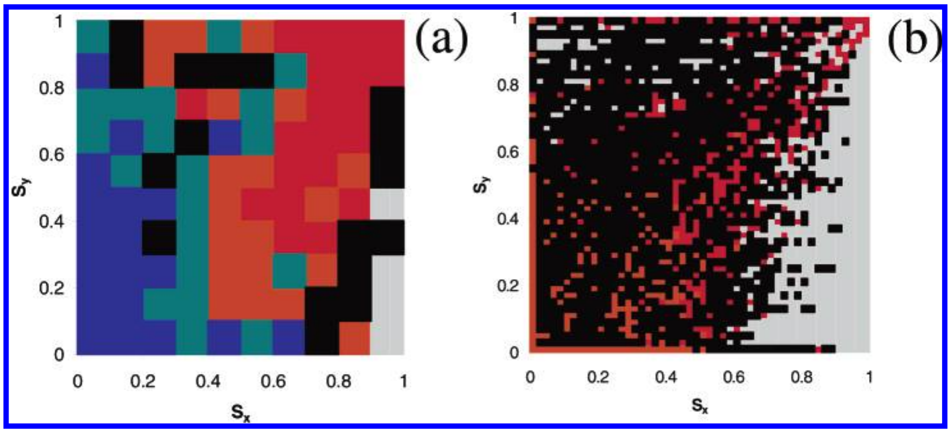

Figure 8. Precision maps for the Scaled Squared Euclidean (S_x) and Forbes (S_y) coefficients obtained using 83 wound healing agents from the MDDR as the active set. Results were taken up to rank 2000, and the whole of the MDDR (less the 83 known agents) was used as the nonactive set. (a) 10 × 10 grid, (b) 50 × 50 grid. Color-coded precision ranges: (red square) $P > 0.8$; (orange square) $0.8 \geq P > 0.6$; (green square) $0.6 \geq P > 0.4$; (blue square) $0.4 \geq P > 0.2$; (black square) $0.2 \geq P > 0.0$; (black square) $P = 0$; (gray square) *neutral* − no compounds recovered.

and unmatched components, as do the recovered-nonactive similarities. Moreover, the matched and unmatched components may be partially related, and the distributions arising from the two measures may be correlated or anticorrelated. None of these factors may be safely neglected in evaluating the enhancement, and this complexity is surely responsible for the enigmatic behavior of data fusion results. Moreover, the examples here suggest that accounting for data fusion results is difficult even after the event; the foundations of a positive result are finely balanced and sensitive to rather subtle effects. Nevertheless the tools that we have described do give an insight into the mechanism, and some justification for good or poor performance can sometimes be offered as a result.

In conclusion, then, our analysis has demonstrated the highly complex interplay of factors that takes place when similarity fusion is used with different types of similarity coefficient; so complex indeed that it is not really surprising that previous studies have been unable to provide unequivocal evidence for the general effectiveness of similarity fusion using multiple similarity coefficients.

A Customized Fusion Rule for Scaled Similarities. In this section we consider the construction of the best possible fusion rule for a given set of results. The precision maps show where the most concentrated regions of actives occur, and, as we have seen, the arrangement of these regions affects the recall obtained by fusion over a single measure and thus the fusion enhancement that may be obtained. The best results are obtained when the integration region matches the arrangement of high precision regions. The question arises whether we can use these maps to develop a more generalized shape of fusion rule—in this case perhaps suited to a particular combination of coefficients. To display results and produce a visible map the precision data has first been coarse-grained into seven categories (but is available to higher definition). By ranking the precision values obtained in each grid square and indexing each grid square with that rank it is possible to produce a trained *precision mask* that defines the best way of accessing the actives to maximize the number retrieved while minimizing the rank. An associated fusion rule can then be constructed that consists of scoring each retrieved structure according to this rank value, which is obtained from its position in the correlation diagram and ultimately the similarity values obtained from the two measures. In this way, structures in the highest precision regions, as defined in the trained mask, are accessed first and those in the lowest precision regions are accessed last.

We call this approach precision-directed fusion (PDF). Initial experiments showed, hardly surprisingly, that in all cases the best results were obtained using a precision mask trained by the same activity class for which searches and fusion were being carried out; the results obtained using a mask trained by a different activity class only occasionally showed significant improvements over single-measure or standard fusion. We would thus suggest that the PDF results for a given class obtained using a mask trained on the same class represent an upperbound to the search performance achievable using a data fusion approach. It is, however, just an upperbound and thus unlikely to be achievable with any practical fusion system: it has been developed primarily to provide further insights into the potential of data fusion techniques since, as the discussion above makes clear, it requires extensive training data that are unlikely to be available in practice (where similarity searching is commonly used at an early stage in a lead discovery program when only limited structure−activity relationship information is available).

With this proviso, some results for similarity fusion using different similarity coefficients are shown in Table 4, which also records values of δ, the difference in recall obtained using the two measures, scaled for comparison with the enhancement by the better single value result

$$\delta = \frac{|R_x - R_y|}{\max(R_x, R_y)} \quad (15)$$

The most significant point about the results in Table 4 is that the PDF results for 50 × 50 masks are always positive, sometimes significantly so. This shows unambiguously that effective data fusion is possible *in principle* for these combinations of coefficients: but the comparatively poor and inconsistent results for the traditional fusion methods show that these methods are unable to make effective use of the extra available information. Overall, the results using a 10 × 10 mask are only a marginal improvement over the traditional methods and are sometimes inferior. This suggests

Table 4. Fusion Enhancement for the Combination of Scaled Similarity Values Obtained Using BCI Strings with 1052 Bits Obtained for the Set of Wound Heal Compounds at Rank 100[a]

	recall				enhancement					
	R_1	R_2	d	$\bar{M}_R$	SUM	MAX	MIN	Comb MNZ	PDF(10)	PDF(50)
T_E	**0.062**	0.053	0.149	0.517	−0.026	−0.062	−0.005	−0.026	−0.021	**0.066**
E_F	**0.053**	0.047	0.103	0.151	0.142	0.097	0.036	*0.164*	0.175	**0.482**
E_R	**0.053**	0.033	0.382	0.503	−0.134	−0.326	−0.047	−0.131	0.175	**0.479**
R_F	0.033	**0.047**	0.311	0.067	0.040	−0.180	0.158	*0.286*	0.239	**0.612**
Y_K	0.048	**0.061**	0.219	0.629	−0.050	−0.202	0.000	−0.050	0.000	**0.197**
T_R	**0.062**	0.033	0.474	0.267	−0.216	−0.438	−0.137	−0.209	−0.007	**0.235**

[a] Each row shows results for a given combination of similarity coefficients. Columns R_1 and R_2 give the single-measure recall values for the first and second similarity coefficient with the best value shown bold, δ is the scaled difference eq 15, $\bar{M}_R$ is the average match ratio eq 8, and the remaining columns give the enhancement, ϵ, using the SUM, MAX, MIN, and CombMNZ rules. The traditional fusion result with the best positive enhancement in each row is shown in italics. The final columns show the enhancement obtained by precision directed fusion (PDF) using a 10 × 10 and 50 × 50 masks trained by the same activity class. The latter are always the best overall fusion results and are shown bold.

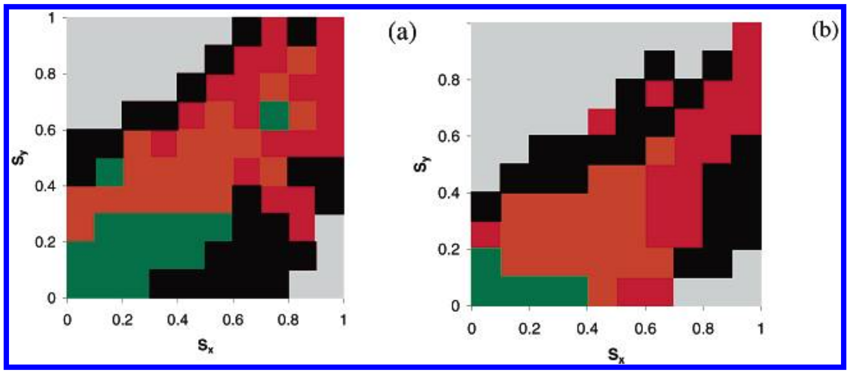

Figure 9. Precision map for the scaled Tanimoto similarities for the 100 reverse transcriptase inhibitors activity class from the MDDR. Obtained using (a) BCI strings with 1052 bits (S_x) and 4096 bits (S_y) and (b) BCI strings with 1052 bits (S_x) Scitegic ECFP4 strings (S_y). Results were taken up to rank 2000, and the whole of the MDDR (less the 100 known agents) was used as the nonactive set. Color-coded precision ranges: (red square) $P > 0.8$; (orange square) $0.8 \geq P > 0.6$; (green square) $0.6 \geq P > 0.4$; (blue square) $0.4 \geq P > 0.2$; (black square) $0.2 \geq P > 0.0$; (black square) $P = 0$; (gray square) *neutral* − no compounds recovered.

that, for these combinations, the optimal fusion rule is a complex arrangement of relatively fine mesh points that cannot be adequately represented even by a 10 × 10 mask. An example of a 50 × 50 precision map is given in Figure 8b and demonstrates this point clearly.

Similarity Fusion Using Different Representations. Similarity fusion can also be carried out using similarity measures based on different representations. Here, we have considered the combination of BCI bit-strings of length 1052 and 4096 using the Tanimoto coefficient to measure similarity. Although we might expect these representations to be strongly correlated, they are based on different dictionaries of substructural fragments, and the precision map in Figure 9a reveals that there is a significant spread in the results and that data fusion may therefore be worth applying. We have also combined BCI bit-strings of length 1052 with Scitegic ECFP4 strings, which are based on circular substructures,[12] as shown in Figure 9b.

Results using the SUM, MAX, and MIN rules are compared with the PDF results in Table 5. We found that in these cases, a 50 × 50 grid produced significantly better results than a 10 × 10 grid (when using a precision mask trained by the same activity class). Enhancements using the traditional fusion rules are, in general, comparatively modest and frequently negative. This is even true of PDF using 10 × 10 grids, where the results are only sometimes an improvement over traditional fusion; however, using 50 × 50 grids always gives the best enhancements and again shows that fusion is possible in principle but difficult in practice.

An analysis of the contribution from unmatched recall for these results is given in Table 6. The zero values obtained for the MIN-rule are consistent with our finding that this rule collects matched values first.[9] Also, of the traditional fusion rules, the MAX-rule always collects the largest number of unmatched results as expected. The PDF results are in most cases roughly equivalent to the MAX- or SUM-rule values but overall show no particularly strong correlation. Since the PDF results always give the best enhancement, the results shown here give the optimal proportion of unmatched values. However, this proportion varies between data sets and does not show any particular trend with diversity (see Table 1).

ANALYSIS OF GROUP FUSION

Group fusion is the combination of similarity lists from several reference compounds belonging to the same activity class. We have found it to be consistently successful, in marked contrast to similarity fusion.[8] In our model, each individual list may be associated with a similarity distribution and the process formally described as a multiple integration with a fusion rule described as a region of hyperspace.[9] To cover this space efficiently we choose maximally diverse sets (generated using a Max-Min algorithm with respect to the Tanimoto coefficient) of the active molecules in each activity class. It is, of course, difficult to visualize the method above three dimensions. However, we can develop precision maps for all pairwise combinations of lists. Thus, for three lists, the combined similarity distribution fills a cube.

Table 5. Fusion Enhancement for the Combination of Scaled Similarity Values Obtained Using the Tanimoto Coefficient at Rank 100 and (a) BCI Strings with 1052 and 4096 Bits and (b) BCI Strings with 1052 and Scitegic ECFP4 Strings[a]

(a) BCI Strings with 1052 and 4096 Bits

	recall				enhancement					
	BCI 1052	BCI4096	d	$\bar{M}_R$	SUM	MAX	MIN	Comb MNZ	PDF(10)	PDF(50)
RTI	0.037	**0.047**	0.208	0.215	−0.047	−0.076	−0.091	−0.057	−0.073	**0.185**
WDH	**0.062**	0.061	0.017	0.249	*0.073*	0.024	0.062	0.071	0.095	**0.239**
ACH	**0.056**	0.049	0.122	0.154	*0.026*	0.01	−0.079	*0.026*	0.02	**0.258**
HIV	0.049	**0.056**	0.123	0.252	0.029	−0.073	0.004	*0.044*	0.025	**0.304**
DAG	0.135	0.121	0.106	0.164	*0.06*	0.037	−0.034	0.023	0.053	**0.072**
PEN	0.23	**0.307**	0.251	0.376	−0.085	−0.226	−0.079	−0.093	0.031	**0.085**

(b) BCI Strings with 1052 and Scitegic ECFP4 Strings

	recall				enhancement					
	BCI 1052	ECFP4	d	$\bar{M}_R$	SUM	MAX	MIN	Comb MNZ	PDF(10)	PDF(50)
RTI	0.037	**0.039**	0.047	0.156	0.069	0.016	0.013	*0.082*	0.056	**0.279**
WDH	0.062	**0.07**	0.115	0.282	−0.019	−0.073	−0.015	−0.023	0.004	**0.126**
ACH	0.056	**0.078**	0.283	0.143	−0.138	−0.184	−0.118	−0.108	−0.025	**0.277**
HIV	0.049	**0.061**	0.196	0.24	−0.013	−0.143	*0.042*	0.008	0.004	**0.356**
DAG	0.135	**0.14**	0.035	0.175	0.051	−0.018	0.051	*0.061*	0.035	**0.076**
PEN	0.23	**0.416**	0.448	0.331	−0.208	−0.408	−0.137	−0.2	−0.03	**0.067**

[a] Each row shows results for a given activity class. The traditional fusion result with the best positive enhancement in each row is shown in italics. The first two columns of results show the single-measure recall with the best value shown in bold. Other columns are as in Table 4.

Table 6. Ratio of Unmatched to Total Recall for the Combination of Scaled Similarity Values Obtained Using BCI Strings with 1052 and 4096 Bits for Each Fusion Rule at Rank 100[a]

activity	BCI 1052: BCI 4096				BCI 1052: ECFP4			
class	SUM	MAX	MIN	PDF(50)	SUM	MAX	MIN	PDF(50)
RTI	0.019	0.042	0.000	0.099	0.091	0.139	0.000	0.086
WDH	0.015	0.079	0.000	0.008	0.011	0.061	0.000	0.006
ACH	0.034	0.109	0.000	0.063	0.041	0.111	0.000	0.114
HIV	0.030	0.128	0.000	0.075	0.015	0.053	0.000	0.110
DAG	0.033	0.111	0.000	0.103	0.027	0.078	0.000	0.031
PEN	0.003	0.017	0.000	0.092	0.005	0.032	0.000	0.178

[a] The final column in each case shows the ratios obtained by precision directed fusion (PDF) using a 50 × 50 mask trained by the same activity class.

Projections of this volume distribution onto the three faces that include the point of maximum similarity (1,1,1) correspond to the three possible pairwise combinations. Averaging these gives a single two-dimensional distribution that can then be converted to a precision map. If $P_{12}(g)$ represents the precision found in square g for the combination of similarities from targets *1* and *2* (see eq 9), then, for targets combined in reverse order, the same density must appear in a cell g', the reflection of cell g in the 2D $y = x$ diagonal; hence, $P_{21}(g') = P_{12}(g)$. However, all fusion rules that we consider here are symmetric to interchange of targets: they give the same result if lists are taken in a different order. Thus, for the SUM, MAX, and MIN rules, cells g and g' are equivalent, and we need to consider only one ordering for each pair of similarities. We therefore need to visit each pairwise combination only once; hence, for three measures an average pairwise precision can be obtained from $(1/3)\{P_{12}(g) + P_{13}(g) + P_{23}(g)\}$, and for m targets this can be generalized to

$$P(g) = (1/m^2 - m)\sum_{j>i}^{m}\sum_{i=1}^{m} P_{ij}(g) \tag{16}$$

A symmetric version could be obtained by summing over all $j \neq i$. We can thus obtain a visible representation corresponding to an average view of what is happening in the higher dimensional space. The Product-Moment Correlation Coefficient is similarly defined only for pair combinations, but by averaging over all pairs we can arrive at a value that may have some significance for the multidimensional problem. In Table 7, we give values of ρ evaluated in this way along with the average pairwise match ratio, obtained in a similar manner using eq 8, and compare these values with the relative improvement index[8] ΔR in the final column. This is the appropriate measure of enhancement for group fusion and compares the group recall R_G with the single measure result represented by the average recall R_{av}.

$$\Delta R = \frac{R_G - R_{av}}{R_{av}} \tag{17}$$

This table shows that average values of the pairwise match ratio are generally very small with the exception of the results for PEN. This activity class is composed of rather similar molecules (Table 1), which also accounts for the low value of ΔR; even a single reference structure is relatively good at finding the other members of the activity class, and group fusion has little scope for further improvement[9] (whereas for the other classes here the ΔR values show a significant increase in performance). Values of ρ are negative or very low. Since the unmatched values predominate numerically, much of the negative contribution is from these. However, the contribution from the matched values although small is also negative in most cases. Negative values of ρ signify that anticorrelation is present and that similarity values tend to concentrate in the top-left and bottom-right quadrants relative to the means. This indicates that even if a compound is judged highly similar to a given reference structure it is likely to be measured as relatively different to another reference structure. The use of maximally diverse subsets of reference structures seems a likely origin and undoubtedly exacerbates this effect. However, negative values of ρ persist when random subsets of reference structures are used—even though these references may well be structurally similar in

Table 7. Comparison of Group Fusion Results for Lists of Length 2000 for the Activity Classes Given in Table 1, Each Using Subsets of 50 Maximally Diverse Structures as Reference Structures[a]

	all recovered				recovered-actives				
activity class	pairs $\bar{M}_R$	all r	unmatched ρ_u	matched ρ_m	pairs $\bar{M}_R$	all r	unmatched ρ_u	matched ρ_m	ΔR
RTI	0.031	−0.23	−0.272	−0.063	0.039	−0.393	−0.438	0.188	6.21
WDH	0.038	−0.283	−0.332	−0.024	0.06	−0.517	−0.586	0.167	4.77
ACH	0.046	−0.2	−0.286	−0.021	0.09	−0.332	−0.436	0.065	2.95
HIV	0.064	−0.219	−0.305	−0.006	0.08	−0.354	−0.441	−0.013	1.69
DAG	0.055	−0.191	−0.28	−0.016	0.083	−0.379	−0.455	−0.04	2.32
PEN	0.302	0.067	−0.318	0.09	0.504	−0.11	−0.426	−0.042	0.33

[a] Results are split into those obtained from the whole comparison lists (all recovered) and those obtained using only the recovered actives in each list (recovered-active). In each case the match ratio for pairs from eq 8 is given as an average $\bar{M}_R$, followed by the Product-Moment Correlation Coefficient for all similarity values considered and for the unmatched and matched components. The final column gives the improvement index ΔR using the SUM-rule at rank 2000.

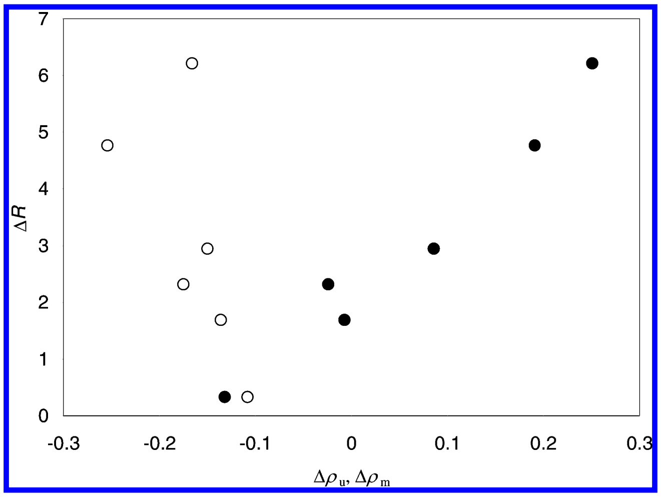

Figure 10. The relative improvement index for each activity class (Table 8) plotted against the difference in the Product-Moment Correlation Coefficient between recovered-active and recovered-nonactive values: $\Delta\rho_u$ ○ unmatched and $\Delta\rho_m$ ● matched.

some way as a result of the similar property principle. Matched values of ρ for the recovered-active comparisons are slightly more positive than those for the recovered-nonactive comparisons, but the converse is true for the unmatched contributions. Figure 10 shows the relative improvement index ΔR plotted against the difference in the Product-Moment Correlation coefficient between recovered-active and recovered-nonactive values. Since unmatched values greatly outnumber the matched values, the plot for unmatched values best reflects the overall results, which are omitted for clarity. Large values of ΔR seem to be associated with positive differences in ρ for the matched values and negative differences in ρ for the unmatched values.When multiple lists are combined we can use a generalized definition of the match ratio, M_R, eq 8, by considering those structures that are matched at least once

$$M_R^G = \frac{\text{no. of structures common to more than one list}}{\text{total no. of different structures in all lists}} \qquad (18)$$

In this case the average $\bar{M}_R^G$ is obtained by visiting all pairwise combinations of the lists. In Table 8 we compare group fusion results in more detail for a single activity class for each of the fusion rules considered. These tables show the results for group fusion using maximally diverse subsets of increasing size m chosen from the activity class. We have again chosen the WDH activity class, partly for continuity with our earlier section on similarity fusion but also because it provides an example of two unusual behaviors. For this

Table 8. Comparison of Group Fusion Results Using the Tanimoto Coefficient and Lists of Length 2000 for the Set of 83 Wound Healing Agents with Maximally Diverse Subsets Size m as Reference Structures[a]

		normalized recall area				
m	$\bar{M}_R^G$	matched	unmatched	total	ΔR(1000)	ΔR(2000)
			(a) SUM-Rule			
5	0.0284	0.0000	0.0724	0.0724	0.030	−0.055
10	0.1433	0.0000	0.0403	0.0403	−0.485	−0.370
20	0.3516	0.1393	0.0083	0.1476	1.060	0.890
30	0.4389	0.2449	0.0504	0.2953	2.219	2.885
40	0.5164	0.4685	0.0146	0.4831	4.537	5.195
50	0.5840	0.5720	0.0000	0.5720	5.438	5.405
60	0.6165	0.5796	0.0000	0.5796	5.696	5.405
70	0.6224	0.5555	0.0000	0.5555	5.438	5.195
80	0.6270	0.4825	0.0000	0.4825	4.408	4.985
			(b) MAX-Rule			
5	0.0284	0.0000	0.0728	0.0728	0.030	−0.055
10	0.1433	0.0000	0.0476	0.0476	−0.485	−0.055
20	0.3516	0.1040	0.0142	0.1182	0.416	0.890
30	0.4389	0.2073	0.0820	0.2893	2.348	3.305
40	0.5164	0.3676	0.0567	0.4244	3.764	4.565
50	0.5840	0.5503	0.0000	0.5503	5.438	5.510
60	0.6165	0.6009	0.0000	0.6009	6.082	5.615
70	0.6224	0.6126	0.0000	0.6126	6.082	5.615
80	0.6270	0.6499	0.0000	0.6499	6.726	5.825
			(c) CombMNZ			
5	0.0284	0.0000	0.0690	0.0690	0.030	−0.055
10	0.1433	0.0000	0.0209	0.0209	−0.871	−0.580
20	0.3516	0.1564	0.0000	0.1564	0.932	0.890
30	0.4389	0.2576	0.0000	0.2576	1.962	1.835
40	0.5164	0.3847	0.0000	0.3847	3.507	4.145
50	0.5840	0.4163	0.0000	0.4163	3.620	4.670
60	0.6165	0.4251	0.0000	0.4251	3.507	5.195
70	0.6224	0.3980	0.0000	0.3980	4.022	4.985
80	0.6270	0.2891	0.0000	0.2891	2.219	3.935
			(d) MIN-Rule			
5	0.0284	0.0000	0.0502	0.0502	−0.614	−0.055
10	0.1433	0.0060	0.0000	0.0060	−1.000	−0.895
20	0.3516	0.0763	0.0000	0.0763	−0.099	0.050
30	0.4389	0.0639	0.0000	0.0639	−0.227	−0.160
40	0.5164	0.1006	0.0000	0.1006	0.288	0.050
50	0.5840	0.0877	0.0000	0.0877	0.159	−0.055
60	0.6165	0.0400	0.0000	0.0400	−0.485	−0.580
70	0.6224	0.0097	0.0000	0.0097	−0.871	−0.895
80	0.6270	0.0000	0.0000	0.0000	−1.000	−1.000

[a] For fusion using (a) SUM-rule, (b) MAX-rule, (c) CombMNZ, and (d) MIN-rule. The match ratio, eq 18, is given as an average $\bar{M}_R^G$, followed by the normalized recall areas, eq 19, at rank 2000. The final columns give the fractional improvement index ΔR at rank 1000 and 2000.

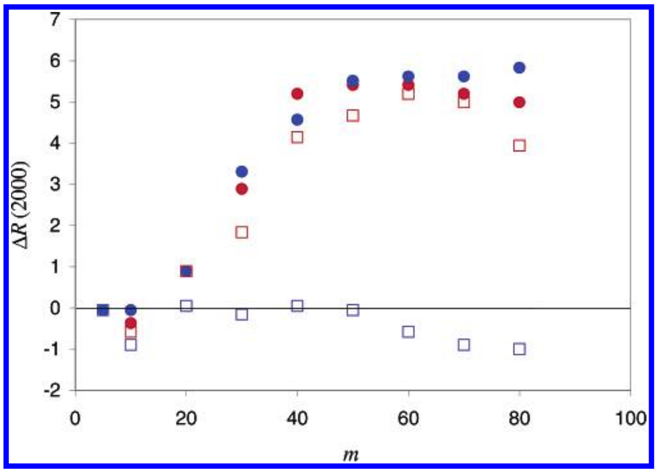

Figure 11. The group recall improvement index Δ*R* for the 83 wound healing agents plotted against the reference set size *m* for (red circle) SUM-rule; (blue circle) MAX-rule; (red square) CombMNZ; (blue square) MIN-rule.

set we found, using the SUM-rule, that the group recall at rank 2000 falls below the average recall when low numbers of targets ($m \leq 10$) are used, leading to a negative improvement index. The recall then passes through a maximum at $m \approx 55$ before falling significantly as *m* approaches the full set size of 83. The improvement index for each fusion rule is plotted against *m* in Figure 11 where these characteristics can be seen for the SUM-rule and the related CombMNZ.

Results for MAX-rule, however, show a less pronounced negative region for $m \leq 10$ and pass through a plateau region at $m \approx 55$ before increasing again as *m* approaches the full set size. Results for the MIN-rule are very poor and even negative over much of the range. A rule that emphasizes the combination of matched values is predictably a poor choice for a system that is highly unmatched.

Because the recall may depend on rank we have summarized the overall performance at all values of rank available with the area under the recall curve normalized by the rank, *n*, at the measurement point

$$A_R(n) = \frac{1}{n}\sum_{r=1}^{n} R(r) \qquad (19)$$

The purpose of this is to encapsulate the overall performance in one recorded value. The normalized recall area is split into matched and unmatched portions, which are also recorded in the table.

In the initial stages, because of the very low match ratios involved, the normalized recall area results show that all of the actives recovered are from unmatched values. An active found in one list is unlikely to be found in another, especially since the reference compounds are chosen to be maximally diverse and thus have a tendency to choose different comparison compounds. The associated normalized precision area (not shown) for these regions is comparatively low so that the unmatched regions are a comparatively dilute source of active compounds. As the number of targets increases, the increasing match ratio shows that the fused list contains more matched compounds. The proportion of unmatched compounds contributing to the total recall falls to zero when $m > 40$ for both the SUM and MAX rules. For CombMNZ the unmatched contribution vanishes much earlier ($m > 10$), since this rule is weighted toward matched compounds, but the results are less successful than those for the SUM-rule.

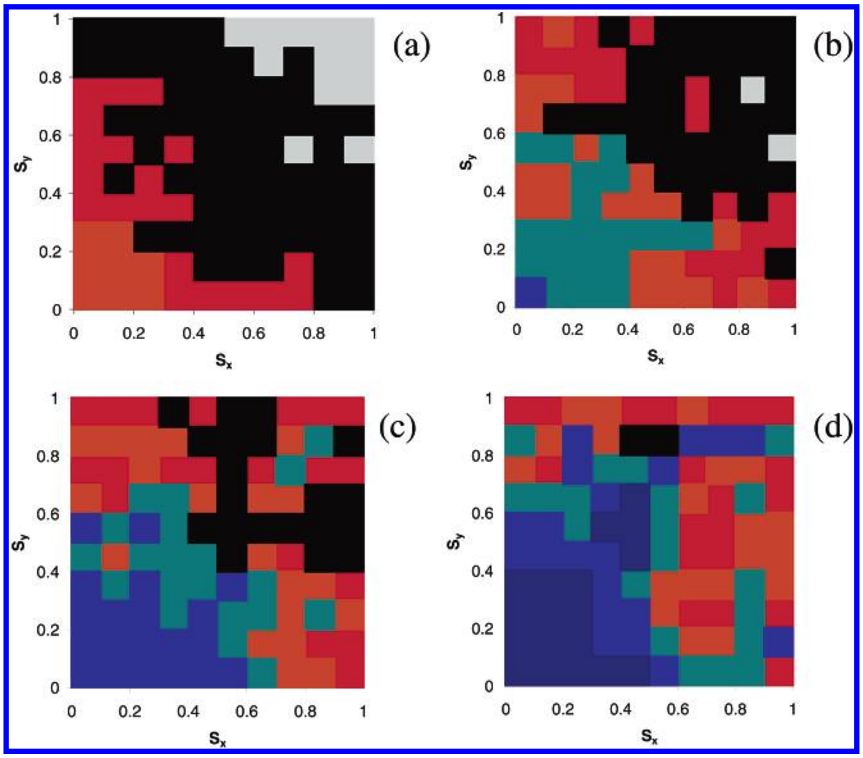

Figure 12. Average pairwise precision maps obtained using the scaled Tanimoto coefficients using 100 reverse transcriptase inhibitors as the active set. Plots show the results obtained using *m* maximally diverse subsets of these as the targets with (a) $m = 40$; (b) $m = 60$; (c) $m = 80$; (d) $m = 100$. Results were taken up to rank 2000, and the whole of the remaining database was used as the nonactive set. Color-coded precision ranges: (red square) $P > 0.8$; (orange square) $0.8 \geq P > 0.6$; (green square) $0.6 \geq P > 0.4$; (blue square) $0.4 \geq P > 0.2$; (black square) $0.2 \geq P > 0.0$; (black square) $P = 0$; (gray square) *neutral* − no compounds recovered.

Thus, in the region where significant improvement is obtained most or all of the contributing values are matched.

The pairwise precision maps for group fusion, as computed from eq 16, show comparable behavior for all of the activity classes studied, but they are quite different from the maps we have seen for similarity fusion. Results for reverse transcriptase inhibitors are shown as an example in Figure 12 for representative values of *m*. For low values of *m* high precision regions appear first near the axes since a high proportion of unmatched values are involved, see Figure 12(a). As more reference structures are used the major regions are seen to be anticorrelated, appearing mainly in the top left and bottom right quadrants as in Figure 12(b). Although the maps shown correspond to combinations taken in a given order, as indicated by eq 16, a degree of symmetry about the $y = x$ diagonal appears naturally in the results, and this was true for all classes examined. Using model bivariate distributions[9] we found that only the MAX-rule could lead to consistent fusion enhancement over the full range of rank for anticorrelated matched values, in contrast to the SUM-rule, which results in negative values at high rank.[9] When a large proportion of the available actives are used as targets, Figure 12(d), the areas of high average precision move to the regions of highest similarity, and the comparison with the MAX-rule region, shown in Figure 4(b), is striking.

Although the precision maps shown in Figure 12 are generated for pairwise combinations of lists, they show that the highest concentrations of active matched compounds appear in regions that correspond closely with the two-dimensional MAX-rule region. We have also seen (Table 8) that group fusion depends wholly on matched values for results when *m* is high. These maps are a projection from multidimensional space onto a two-dimensional window, but the appearance of these distributions seems to underlie the superiority of the MAX-rule over the SUM-rule when group fusion is used. Furthermore, since results using the MAX-rule do not present an optimal value of *m* (Figure 11), it would appear that the reason for the falloff in performance associated with the SUM-rule could be linked with the relative mismatch of the SUM-rule region with distributions that favor the MAX-rule when *m* is high.

CONCLUSIONS

We have applied some recently developed methods[9] to the examination of data fusion methods in similarity-based virtual screening. Our methods involve representing typical data fusion rules by multiple integration of similarity frequency distributions and help to explain why fusion is successful in some cases and not in others. As a predictive tool our methods may be less helpful because of the complexity of the problem, since the fusion of even just two lists of similarity values depends on eight distinct distributions (the recovered-active and recovered-nonactive distributions of both lists for both matched and unmatched compounds). However, if a precision map for a pair of similarity lists yields a pattern closely related to the integration region associated with a given fusion rule and if the match ratio is high, then there is a good chance that fusion by that rule will be successful. For low match ratios, many values lie on the axes of the precision map, and the relative efficacy of different fusion regions is more difficult to judge. The extension of this criterion to higher dimensions, and thus the combination of several lists, is problematical, but the use of pairwise average precision maps appears to be useful for descriptive purposes.

We have examined similarity fusion, using both different similarity coefficients and different representations, in terms of the contributions from matched and unmatched components and also in terms of precision maps that graphically display the richest seams of recovered-active similarity values. While it is clear that close compatibility of a region associated with a given rule and the arrangement of high-precision regions can lead to positive fusion enhancement, we have also found that, in practice, the success or otherwise of data fusion can be based on rather subtle distinctions in the arrangement of high precision regions. In many cases, including our choices of different representations, the conventional fusion rules often make a rather poor match to the high precision regions, resulting in poor fusion results. Group fusion, in contrast, is consistently successful. We have shown that although the data sets used here are highly unmatched if pairs of lists are taken, the fusion of many such lists leads to recall that is the resultant of partially matched values, i.e., those that are matched in some lists. Precision maps obtained by averaging over all pairwise combinations of these lists show that partially matched values associated with active compounds become concentrated in high similarity regions as the number of targets increases. Moreover, these regions are qualitatively very similar to the two-dimensional MAX-region. Although this evidence is only a two-dimensional projection of a high dimensional problem, the persistent appearance of this pattern for each of the active classes that we have examined seems to lie at the root of the superiority of MAX over the SUM-rule for group fusion that has been observed previously.[8]

In conclusion, while we believe that our theoretical and experimental work provide a useful interpretive tool for the analysis of data fusion results, the complexities that we have identified in the fusion process means that it will be difficult to develop fusion methods that can be expected consistently to outperform individual similarity searches. Our use of precision maps to rank each grid square according to its retrieval effectiveness provides, for the first time, an estimate of the maximum possible benefit that might be achieved by similarity fusion in searches of a specific activity class using a given similarity measure. Significant improvements over standard fusion rules can be obtained, especially at high rank; however, in practice this upperbound level of performance is extremely unlikely without considerable amounts of training data. Moreover, if such amounts were available, then it is arguable that alternative types of machine learning should be employed (such as Bayesian classifiers, binary kernel discrimination, or substructural analysis, see, e.g., refs 18−20). These seek to model explicitly the characteristics of active and inactive molecules and thus provide a much more detailed analysis of the training data than does similarity searching, which is based on one or a few known actives.

ACKNOWLEDGMENT

We thank the following: Pfizer Global Research and Development for funding; Alexander Alex for helpful discussions and guidance; Jérôme Hert for extracting the Scitegic ECFP-4 fingerprints and making them available to us; the referees for their comments on this manuscript; Digital Chemistry 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.

REFERENCES AND NOTES

(1) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: a Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem*. **1999**, *42*, 5100−5109.

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

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

(4) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(5) 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.

(6) 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.

(7) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Vvirtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(8) Whittle, M.; Gillet, V. J.; Willet, P.; Loesel, J.; Alexander, A. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest-Neighbour Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.

(9) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Theoretical Model. *J. Chem. Inf. Model.* **2006**, *46*, 2193−2205.

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

(11) The BCI software is available from Digital Chemistry Ltd. at URL http://www.digitalchemistry.co.uk.

(12) The Pipeline Pilot software is available from Scitegic Inc. at URL http://www.scitegic.com.

(13) 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.

(14) 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.

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

(16) Beitzel, S. M.; Jensen, E. C.; Chowdhury, A.; Grossman, D.; Goharian, N.; Frieder, O. Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. *J. Am. Soc. Inf. Sci. Tech*. **2004**, *55*, 859−868.

(17) 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.

(18) 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.

(19) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naive Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.

(20) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model*. **2005**, *45*, 1122−1133.

CI0496144