A Machine Learning Approach to Weighting Schemes in the Data Fusion of Similarity Coefficients

Jenny Chen,[†] John Holliday,*[,†] and John Bradshaw[‡]

Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, U.K., and Daylight Chemical Information Systems Inc., 120 Vantis, Suite 550, Aliso Viejo, California 92656

Received August 20, 2008

The application of data fusion techniques for combining the results of similarity searches of chemical databases has been shown to improve search performance. When used to combine the results of searches using different similarity coefficients, the optimum combination is dependent on the size, in terms of substructural fragments present, of the molecules being compared. This paper describes preliminary simulation tests which aim to automatically deduce, using machine learning techniques, the optimum combination of similarity coefficient which may be combined using data fusion for a given class of active compounds.

INTRODUCTION

Calculations of molecular similarity and dissimilarity have been widely used in the chemoinformatics field in areas[1−5] such as property prediction, virtual screening, database clustering, similarity searching, compound selection, and synthesis design. The technique involves the determination of an appropriate measure of correlation (coefficient) between linear representations of the two molecules in question. The representations used often take the form of a binary fingerprint, a set of binary elements either describing the presence or absence of members of a set of predefined substructural fragments or derived algorithmically using a hashing technique.[6]

Several coefficients have been described in previous studies,[7−12] and these fall into three main classes: association coefficients, correlation coefficients, and distance coefficients. Most of these can be used on binary or nonbinary data, and, in the case of binary data, they generally have values in the range zero to unity where zero indicates no correlation and unity indicates total correlation. In the case of an association coefficient such as the Tanimoto coefficient, for example, a value of zero would indicate that no common features occur in the two binary representations and unity would indicate that they were identical. Distance coefficients, the Mean Manhattan for example, would produce a complementary range of values where zero would indicate identity and unity would indicate no common features. When used for nonbinary data, these ranges may be quite different; however, we will concentrate on their application to binary data in this paper.

These coefficients can be used to search a database of such representations, using a query representation, and the resultant values are sorted to produce a ranked list. Those items, chemical compounds in this instance, at the top of the list are, therefore, more similar to the query compound than those lower down. When the ranked lists produced by different coefficients are compared, using the same query and database, there are varying degrees of correlation. In some cases the rankings are exactly the same, or monotonic, indicating that the coefficients would have the same effect when used in a similarity search of this type; whereas, in other cases, the lists produced are quite different, often very different, with little correlation between the rankings at all. The ranking is a function of both the representation and the coefficient. Unfortunately there is no independent measure of compound similarity to check which combination of representation and coefficient is “best”. Indeed, as has been noted,[13] the choice of features to represent the object is arbitrary, so “... the representation of an object as a collection of features is viewed as a product of a prior process of extraction and compilation...”.

Often the reason for carrying out such a search is to find compounds which would exhibit similar properties to a (set of) target compound(s), where there is no direct way to parametrize the compounds according to the desired property. In order to make the transition from a structure based descriptor to a property use is made of the similar property principal,[14] which states that compounds which are similar in structure are more likely to exhibit similar properties and, therefore, when a bioactive compound is used as a query in a similarity search, the set of compounds retrieved should have a high proportion of compounds exhibiting the same activity. If this is so, then the relative performance of each coefficient-representation combination, when used in a similarity search against an active target, is directly related to the proportion of actives retrieved. Furthermore, it may be possible to increase this performance by combining the ranked lists produced by more than one coefficient using data fusion techniques.

Previous studies used this philosophy by first identifying a subset of complementary coefficients,[8] those whose ranked lists were not monotonic or near monotonic, and then combining these using a data fusion methodology[7,9,15,16] with several aims: first, to see which coefficients can be combined to improve retrieval performance; second, to see what is the

* Corresponding author e-mail: j.d.holliday@sheffield.ac.uk.
[†] University of Sheffield.
[‡] Daylight Chemical Information Systems Inc. (until December 2006).

10.1021/ci800292d CCC: $40.75
Published on Web 01/09/2009

optimum number of coefficients to combine; and finally, to see whether a single optimum combination could be identified.

The results of these and further studies indicated that the optimum combination is highly dependent on the nature of the structures in question, in particular the size in terms of the number of substructural fragments represented in a fingerprint. This related both to the query itself and to the database compounds of interest, those of the same active class. However, the size distribution of active classes can vary considerably, and it is difficult, therefore, to identify an optimum combination of coefficients for each one.

Using several bioactive classes, this paper describes the nature of the relationship between compound size (hereafter synonymous with the number of bits set in the fingerprint representation) and the performance of similarity coefficients both individually and in combination. The paper then describes preliminary methods which, using a simulation approach to machine learning techniques, can be used to select the optimum combination of coefficients to combine in a search for compounds of a given active class.

RELATIONSHIP BETWEEN MOLECULAR SIZE AND SIMILARITY

The relationship between the coefficient applied to a similarity search and the size distribution of the highest-ranking compounds retrieved has been discussed in the literature.[10−12,17−20] In addition, when used for the purpose of selecting a diverse set of compounds from a large data set, different coefficients will produce quite different size distributions for the sets of compounds obtained. The Squared Euclidean coefficient, for example, tends to select larger molecules, when used for diversity selection, than the Tanimoto coefficient. Clearly, this is not an ideal situation and has been addressed by Fligner and co-workers with the derivation of a modified version of the Tanimoto coefficient.[18]

To illustrate the relationship between coefficient and size for similarity searching, we used a selection of 13 similarity coefficients (Table 1) which have been shown to retrieve complementary sets of compounds and applied them to similarity searches of the MDL Drug Data Report database (MDDR)[21] using Daylight fingerprints[22] as the binary representation. The database was first divided into twenty equally populated partitions based on the size of the compounds. The first set then contained compounds with between 0 and 132 bits set, the second with between 133 and 161 bits set, and so on, up to the final set which contained compounds with more than 483 bits set. One query compound was selected from each partition such that its size was as close to the middle of the partition as possible. Searches were carried out using each of the twenty compounds against each of the twenty partitions, using all 13 coefficients. The best performing coefficient, in terms of proportion of nearest 500 neighbors which exhibit the same activity as the query, for each of the 400 combinations is indicated in Figure 1.

The results illustrate some important features. First, that the relative performance is clearly dependent on the combination of the size of the query compound and the size of compounds in the data set, with the Russell/Rao coefficient being the obvious choice for data sets of larger compounds and the Forbes or Simple Match being more appropriate for data sets of smaller compounds. Second, that not only is the Tanimoto the predominant coefficient in terms of spread across many size ranges but also it shows a clear like-for-like tendency, accounting for most of the cells in and around the principal diagonal. Finally, studies have shown[8,9] that data fusion is more successful when the coefficients used in combination are known to perform well on their own. The results of this exercise seem to indicate, then, that there are only a handful of obvious good performers and that the original 13 coefficients can perhaps be reduced to as few as four, those being the Russell/Rao, the Tanimoto, the Simple Match, and the Forbes coefficients.

In a series of studies, we looked at a selection of 20 active classes in detail to examine the effect the chosen coefficient had on retrieval in different size-based partitions of one-half of the MDDR database (50,000 compounds). Searches were carried out using all 13 coefficients and using a variety of compounds of different size for each class. The data set was interrogated, and the results separated out into ten size-based partitions as exemplified in Tables 2−5. The retrieval rate for each partition is given as the percentage of actives retrieved in the top 500 nearest neighbors compared to the actual number of database actives in that size range. Table 2, for instance, shows the results for a search using an antihypertensive with a size of 99 bits. Each row of the table indicates the performance within a partition of the database, e.g. 0 to 100 bits in the first row, for each coefficient.

Again, it is obvious from the results that coefficients have a tendency to perform well in certain size ranges and not in others. In Tables 2−5, the best performing coefficient for each partition is indicated in bold. Table 2 clearly shows that the Simple Match performs well when the partition represents small compounds and Russell/Rao performs well for the larger compounds. Indeed, it is often the case that certain partitions show no retrieval at all for many size ranges. The Simple Match, for example, in Table 2 retrieves no compound larger than 200 bits, and yet these are well represented by the Russell/Rao. One might expect, therefore, that the Simple Match and Russell/Rao, when used in combination, would result in better retrieval coverage and possibly improved performance. Searches using the larger query antihypertensive of Table 3, however, might benefit from a combination of the Forbes, Tanimoto, and Russell/Rao coefficients.

Similar patterns are observed for other active classes. Tables 4 and 5 show the results of searches for hypolipidemics using queries of size 200 bits and 491 bits, respectively. The relative performance of the different coefficients is again obvious, but the choice of best combination is quite different. Table 4 would suggest the Simple Match, Baroni, Tanimoto, and Russell/Rao in combination for a query of this class and of this size, whereas Table 5 would suggest the Forbes, Simple Match, Baroni, and Russell/Rao for the larger query. However, once we have identified the subset of complementary coefficients, how would we apply them in a data fusion strategy? Clearly, some of the coefficients have better coverage across size ranges than others, so it might be appropriate to introduce a weighting scheme based on their coverage. Also, do we limit ourselves to just two, three, or four coefficients in combination, as previous studies[8,9] have shown this to be the optimum methodology?

Table 1. Thirteen Similarity Coefficients[a]

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

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

WEIGHTED COEFFICIENT COMBINATIONS

If we are to combine similarity coefficients using data fusion techniques, we have learned from previous studies that the optimum number of coefficients used should not exceed four and should ideally be around two or three. It would help, therefore, if we could reduce the number of options to about four complementary coefficients, the obvious ones being those identified in Figure 1: the Russell/Rao, the Tanimoto, the Simple Match, and the Forbes coefficients. By selecting these four alone, the results shown in Table 3 now become those of Table 6, which includes the number of actives in each size range and, in parentheses, the number of actives retrieved. It might be expected that increased coverage and hence improved retrieval would result if a search of a similar sized active of this type were to be conducted using a combination of the coefficients which are identified as being best performers. However, as discussed above, would it be more appropriate to weight the coefficients accordingly? One weighting scheme, for example, of Table 6 would appear to be 4:0:3:3 (Russell/Rao:Tanimoto:Simple Match:Forbes), since the Forbes is the best performing coefficient in three size ranges, as is the Tanimoto, and the Russel/Rao is best performer in four. Other schemes might, however, be more appropriate.

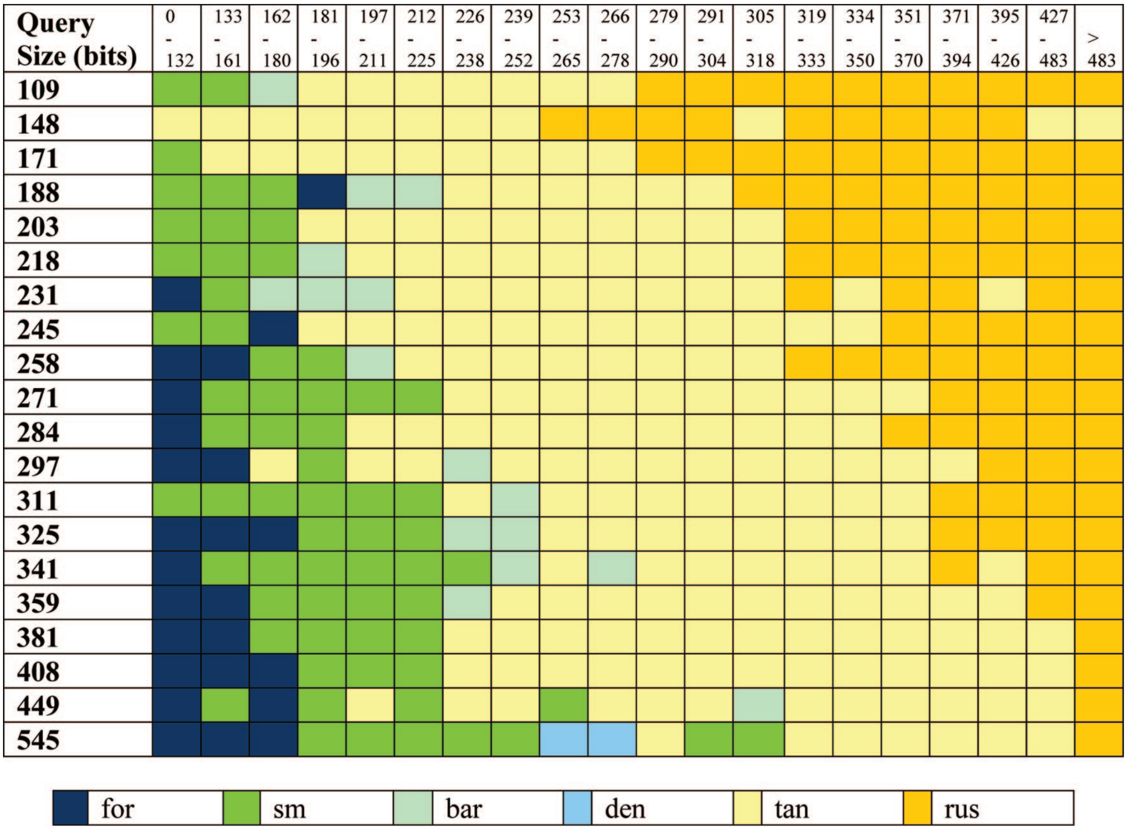
TURBO SIMILARITY SEARCH

Recently, Hert and co-workers[23,24] devised a similarity search methodology called Turbo Similarity Search (TSS), a group fusion algorithm which is applicable when only a single reference structure is available. Turbo similarity searching is based on two notions. First, that combining the results of similarity searches using multiple reference structures, using data fusion methods, is likely to be more effective than searches using a single reference structure. Second, that the nearest neighbors of a reference structure are likely to be similarly active since they share many common chemical features and that, if we assume that they are indeed active, then we can use these nearest neighbors to represent the multiple reference structures. The overall search strategy is summarized in Figure 2.

It would be possible, therefore, to train the weights for each combination of active class and size range using a single known active as our starting point. If more similar-sized actives of this class are known, they can be used to further enhance the training.

SYSTEMATIC WEIGHT SELECTION

Active compounds from between three and four size ranges from eleven active classes, identified by Hert et al.,[25] were used to train the weights of the four coefficients Russell/Rao, Tanimoto, Simple Match, and Forbes. All possible combinations of the weights for the four coefficients were explored systematically in regulated steps, with the sum of all weights always being unity. For example, using incremental steps of 0.2, the first combination (in the ratios Russell/Rao:Tanimoto:Simple Match:Forbes) would be 1.0:0.0:0.0:0.0, the second would be 0.8:0.2:0.0:0.0, and so on. The final two combinations would then be 0.0:0.0:0.2:0.8 and 0.0:0.0:0.0:1.0.



Query Size (bits)	0 - 132	133 - 161	162 - 180	181 - 196	197 - 211	212 - 225	226 - 238	239 - 252	253 - 265	266 - 278	279 - 290	291 - 304	305 - 318	319 - 333	334 - 350	351 - 370	371 - 394	395 - 426	427 - 483	> 483
109	sm	sm	bar	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus	rus	rus	rus
148	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	tan	rus	rus	rus	rus	rus	tan	tan
171	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus	rus	rus	rus
188	sm	sm	sm	for	bar	bar	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus	rus
203	sm	sm	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus
218	sm	sm	sm	bar	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus
231	for	sm	bar	bar	bar	tan	tan	tan	tan	tan	tan	tan	tan	rus	tan	rus	rus	tan	rus	rus
245	sm	sm	for	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus
258	for	for	sm	sm	bar	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus	rus	rus
271	for	sm	sm	sm	sm	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus
284	for	sm	sm	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus	rus
297	for	for	tan	sm	tan	tan	bar	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus
311	sm	sm	sm	sm	sm	sm	tan	bar	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus
325	for	for	for	sm	sm	sm	bar	bar	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus	rus	rus
341	for	sm	sm	sm	sm	sm	sm	bar	tan	bar	tan	tan	tan	tan	tan	tan	rus	tan	rus	rus
359	for	for	sm	sm	sm	sm	bar	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus	rus
381	for	for	sm	sm	sm	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus
408	for	for	for	sm	sm	sm	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	tan	rus
449	for	sm	for	sm	tan	sm	tan	tan	sm	tan	tan	tan	bar	tan	tan	tan	tan	tan	tan	rus
545	for	for	for	sm	sm	sm	sm	sm	den	den	tan	sm	sm	tan	tan	tan	tan	tan	tan	rus

■	for	■	sm	■	bar	■	den	■	tan	■	rus

Figure 1. Relationship between best-performing coefficient and retrieval size. The best-performing coefficient is indicated for each combination of query size and database partition size.

Table 2. Percentage Retrieval Rate for 10 Database Size Ranges and 13 Coefficients Using Daylight Fingerprints with the MDDR Database, for Query Size 99, Antihypertensive Class[a]

	Tan	Rus	Sm	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
0−100	61.5	5.1	**97.4**	87.2	35.9	30.8	92.3	35.9	10.3	46.2	71.8	46.2	64.1
101−150	46.6	1.1	**81.5**	58.2	33.9	16.4	58.7	33.3	1.1	37.0	39.7	36.5	41.8
151−200	29.0	0.8	24.3	**33.1**	21.9	9.3	31.7	21.9	0.5	24.3	18.6	23.5	25.4
201−250	10.1	2.9	0.0	4.5	**10.8**	6.1	3.7	**10.8**	2.9	10.3	6.4	10.4	10.1
251−300	5.8	**9.1**	0.0	4.8	8.2	8.4	4.7	8.3	**9.1**	7.5	6.5	7.6	6.7
301−350	9.6	**13.5**	0.0	2.6	9.6	10.3	1.9	9.6	**13.5**	9.6	9.6	9.6	9.6
351−400	7.3	**16.5**	0.0	0.0	9.6	10.6	0.0	9.6	16.2	9.4	9.5	9.4	9.3
401−450	0.2	**14.0**	0.0	0.0	4.9	7.3	0.0	4.9	12.9	4.9	4.9	4.9	3.7
451−500	0.0	**16.7**	0.0	0.0	7.1	8.7	0.0	7.5	16.3	4.8	7.5	5.5	0.4
above 500	0.0	**31.1**	0.0	0.0	1.9	6.6	0.0	1.9	30.2	0.0	5.7	0.0	0.0.

[a] The best-performing coefficient for each database partition is shown in bold. The Simple Match, for example, identifies 81.5% of all antihypertensives in the 151−200 size range.

Table 3. Percentage Retrieval Rate for 10 Size Ranges and 13 Coefficients Using Daylight Fingerprints with the MDDR Database for Query Size 300, Antihypertensive Class

	Tan	Rus	Sm	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
0−100	0.0	0.0	20.0	0.0	0.0	10.3	**40.0**	0.0	**40.0**	12.2	38.5	0.0	5.1
101−150	1.6	0.0	23.3	9.5	7.9	15.3	**34.4**	7.9	32.3	26.0	24.3	12.2	16.4
151−200	19.4	1.9	35.0	26.0	22.4	24.3	**35.2**	22.4	28.7	29.0	28.4	25.7	27.0
201−250	**29.8**	14.1	28.8	31.1	28.3	27.6	26.1	28.3	23.7	29.0	26.8	29.0	29.1
251−300	**25.9**	19.9	21.2	25.5	25.1	23.3	15.7	25.1	12.4	24.7	21.1	24.7	24.0
301−350	**26.8**	24.7	20.4	24.9	25.8	24.4	15.3	25.8	17.9	24.5	21.1	24.5	23.7
351−400	19.9	**22.6**	12.7	15.9	19.1	17.8	10.0	19.2	13.7	16.9	13.7	16.9	15.4
401−450	8.2	**17.7**	3.9	6.1	8.2	7.7	2.4	8.2	7.3	7.3	6.0	7.3	6.1
451−500	5.6	**25.4**	0.0	2.8	6.0	5.6	0.0	6.0	6.7	4.4	2.8	4.4	2.8
above 500	3.8	**45.3**	0.0	0.9	4.7	4.7	0.0	4.7	9.4	1.9	0.9	1.9	0.9

For each combination, the TSS methodology was adapted using three different approaches as follows (refer to Figure 2):

TurboModal (TMod): The fingerprints representing the R nearest neighbors were combined to produce a modal fingerprint, a single fingerprint in which each bit is set to

Table 4. Percentage Retrieval Rate for 10 Size Ranges and 13 Coefficients Using Daylight Fingerprints with the MDDR Database for Query Size 200, Hypolipidemic Class

	Tan	Rus	Sm	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
0−100	20.6	0.0	**89.7**	44.1	33.8	39.7	70.6	33.8	50.0	39.7	60.3	39.7	42.6
101−150	22.6	5.6	**40.5**	30.8	22.6	21.5	39.5	22.6	22.6	26.2	32.3	26.2	28.7
151−200	13.4	4.9	14.1	**16.2**	13.0	11.8	15.0	13.0	7.2	13.4	14.6	13.4	14.4
201−250	**10.2**	4.8	5.3	**10.2**	9.4	8.5	7.3	9.4	3.6	9.4	8.7	9.4	**10.2**
251−300	**10.4**	9.1	1.5	9.8	**10.4**	10.1	3.5	**10.4**	7.3	10.1	8.3	10.1	10.1
301−350	17.3	**21.4**	0.0	5.0	18.1	17.8	0.0	18.1	18.7	14.5	3.3	15.3	8.6
351−400	5.0	**23.9**	0.0	1.0	7.5	12.9	0.0	7.5	21.9	3.0	2.5	3.5	2.5
401−450	0.0	**18.9**	0.0	0.0	0.0	3.3	0.0	0.0	16.7	0.0	0.0	0.0	0.0
451−500	0.0	**39.6**	0.0	0.0	0.0	0.0	0.0	0.0	**39.6**	0.0	0.0	0.0	0.0
above 500	0.0	**46.4**	0.0	0.0	0.0	0.0	0.0	0.0	42.9	0.0	0.0	0.0	0.0

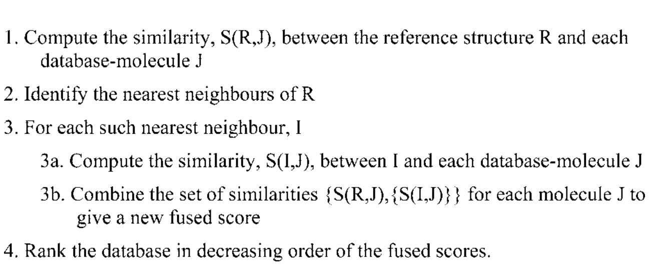
Table 5. Percentage Retrieval Rate for 10 Size Ranges and 13 Coefficients Using Daylight Fingerprints with the MDDR Database for Query Size 491, Hypolipidemic Class

	Tan	Rus	Sm	Bar	Cos	Ku2	For	Fos	Sim	Pea	Yul	Sti	Den
0−100	0.0	0.0	0.0	0.0	0.0	1.5	**14.7**	0.0	**14.7**	0.0	10.3	0.0	0.0
101−150	0.0	0.0	0.0	0.0	0.0	0.0	**7.2**	0.0	**7.2**	0.0	3.1	0.0	0.0
151−200	0.0	0.0	3.2	0.5	0.7	2.8	**5.1**	0.7	**5.1**	2.3	4.4	2.3	2.8
201−250	2.2	0.0	**6.1**	2.4	2.4	4.1	3.4	2.4	3.4	4.8	4.6	4.8	5.1
251−300	2.0	0.5	2.3	**2.8**	2.3	2.3	0.8	2.3	0.8	2.3	1.3	2.3	2.3
301−350	3.6	1.1	2.5	**3.9**	3.3	2.5	0.0	3.3	0.0	3.1	1.1	3.1	2.8
351−400	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
401−450	2.2	**3.3**	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
451−500	4.2	**8.3**	2.1	4.2	4.2	2.1	4.2	4.2	4.2	2.1	4.1	2.1	2.1
above 500	3.6	**46.4**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 6. Four Best Performing Coefficients Showing the Percentage of Actives Retrieved in Each Size Range[a]

	no. of actives	Rus	Tan	Sm	For
0 −100	30	0.0 (0)	0.0 (0)	20.0 (6)	**40.0 (12)**
101−150	189	0.0 (0)	1.6 (3)	23.3 (44)	**34.4 (65)**
151−200	366	1.9 (7)	19.4 (71)	35.0 (128)	**35.2 (129)**
201−250	594	14.1 (84)	**29.8 (177)**	28.8 (171)	26.1 (155)
251−300	772	19.9 (154)	**25.9 (200)**	21.2 (164)	15.7 (121)
301−350	1026	24.7 (253)	**26.8 (275)**	20.4 (209)	15.3 (157)
351−400	853	**22.6 (193)**	19.9 (170)	12.7 (108)	10.0 (85)
401−450	587	**17.7 (104)**	8.2 (48)	3.9 (23)	2.4 (14)
451−500	252	**25.4 (64)**	5.6 (14)	0.0 (0)	0.0 (0)
above 500	106	**45.3 (48)**	3.8 (4)	0.0 (0)	0.0 (0)

[a] The number of actives retrieved are shown in parentheses. The best performing coefficients are shown in bold. The query size is 300 bits, and the active type is antihypertensive.



1. Compute the similarity, S(R,J), between the reference structure R and each database-molecule J
2. Identify the nearest neighbours of R
3. For each such nearest neighbour, I
    - 3a. Compute the similarity, S(I,J), between I and each database-molecule J
    - 3b. Combine the set of similarities {S(R,J),{S(I,J)}} for each molecule J to give a new fused score
4. Rank the database in decreasing order of the fused scores.

Figure 2. Turbo similarity search.

'1' only when the respective bits set to '1' in the nearest neighbor fingerprints exceed a threshold value (60% in this case).

TurboMAX (TMAX): The method for combining the similarity values (3b of Figure 2) uses the maximum similarity value as the new fused score.

TurboSUM (TSUM): The method for combining the similarity values (3b of Figure 2) uses the sum of the similarity value as the new fused score.

Table 7. TurboModal Training Results for Five Example Classes Showing the Query Size, the Number of Actives Retrieved by TurboModal, Tanimoto, and Turbo Tanimoto, and the Improvement Rates of the TurboModal over Tanimoto and Turbo Tanimoto

size	TMod	Tan	TSS	Imp.Tan	Imp.TSS
		Renin Inhibitor			
90	484	422	422	14.7	14.7
115	509	461	484	10.4	5.2
125	520	520	520	0.0	0.0
125	519	492	495	5.5	4.8
135	520	505	507	3.0	2.6
		Thrombin Inhibitor			
91	93	37	37	151.4	151.4
115	242	213	213	13.6	13.6
125	214	214	220	0.0	−2.7
126	181	181	215	0.0	−15.8
126	258	257	258	0.4	0.0
		5HT Reuptake Inhibitor			
90	68	56	63	21.4	7.9
90	47	43	46	9.3	2.2
91	68	50	50	36.0	36.0
115	56	37	37	51.4	51.4
139	73	65	65	12.3	12.3
		5HT1A Agonist			
90	186	150	166	24.0	12.1
105	177	132	140	34.1	26.3
105	92	34	78	170.6	17.9
115	198	126	142	57.1	39.4
138	166	82	118	102.4	40.7
		Cyclooxygenase Inhibitor			
90	72	71	71	1.4	1.4
90	57	37	37	54.1	54.1
91	69	63	64	9.5	7.8
115	64	59	59	8.5	8.5
135	75	54	54	38.9	38.9

In all cases, the number of nearest neighbors is set to 100. The training stage used a data set of 50,000 compounds, half of the MDDR database, and Turbo searching continued until

Table 8. TurboModal Testing Results for Five Example Classes Showing the Query Size, the Weightings Used, the Number of Actives Retrieved by TurboModal, Tanimoto, and Turbo Tanimoto, and the Improvement Rates of the TurboModal over Tanimoto and Turbo Tanimoto

size	Rus	Tan	Sm	For	TMod	Tan	TSS	Imp.Tan	Imp.TSS
				Renin Inhibitor					
90	0.6	0.4	0.0	0.0	507	447	779	13.4	12.9
115	0.6	0.4	0.0	0.0	513	106	513	1.4	0.0
125	0.5	0.5	0.0	0.0	529	504	533	5.0	−0.8
135	0.5	0.5	0.0	0.0	522	471	477	10.8	9.4
				Thrombin Inhibitor					
89	0.6	0.4	0.0	0.0	106	40	40	165.0	165.0
115	0.6	0.4	0.0	0.0	193	56	56	244.6	244.6
125	0.6	0.4	0.0	0.0	235	204	204	15.2	15.2
135	0.6	0.4	0.0	0.0	209	172	173	21.5	20.8
				5HT Reuptake Inhibitor					
90	0.0	1.0	0.0	0.0	44	44	44	0.0	0.0
91	0.0	0.5	0.5	0.0	63	53	53	18.9	18.9
115	0.4	0.2	0.4	0.0	35	38	43	−7.9	−18.6
134	0.1	0.4	0.5	0.0	18	13	13	38.5	38.5
				5HT1A Agonist					
90	0.0	0.5	0.5	0.0	116	85	105	36.5	10.5
105	0.0	0.5	0.0	0.5	154	62	141	148.4	9.2
115	0.0	0.5	0.0	0.5	180	133	163	35.5	10.4
135	0.0	0.5	0.5	0.0	167	131	157	27.5	6.4
				Cyclooxygenase Inhibitor					
90	0.2	0.2	0.3	0.3	23	21	26	9.5	−11.5
90	0.0	1.0	0.0	0.0	67	67	67	0.0	0.0
115	0.3	0.1	0.3	0.3	60	56	56	7.1	7.1
134	0.1	0.4	0.0	0.5	37	50	55	−26.0	−32.7

the retrieval performance, given by the number of actives in the top 5% of the ranked data set compounds, showed no improvement. The result of this operation is a set of weights which represents the optimum combination of weights for the four coefficients for retrieving actives of the reference structure's class and size range. BCI[26] fingerprints were used to represent the database compounds.

The fusion of the weights themselves is rank-based. For each compound, the product of the respective weighting and the inverse rank position is summed across all four coefficients. A final ranking is produced by sorting these summed ranks in reverse order. In the following experiments, the best performing weights are compared with an equivalent Tanimoto search and a Turbo Similarity Search using the same Turbo methodology but with just the Tanimoto coefficient.

TURBOMODAL

TurboModal training was carried out using incremental steps of 0.1 for the weighted combinations of the four coefficients. A selection of five actives was chosen from each of the 11 active classes investigated, and these covered between three and four size ranges, given by the bit density of the fingerprint. The results of the best combination for five of the classes are given in Table 7, which indicates the size of the reference compound (number of bits set), the number of actives retrieved by the best combination (TMod), the number of actives retrieved by the Tanimoto alone (Tan), and the number of actives retrieved by a standard TSS search using the Tanimoto coefficient (TSS). The last two columns indicate the relative performance of TurboModal against Tan and TSS, in terms of percentage improvement rate. Zero indicates no improvement, and 100 indicates that twice as many actives have been retrieved.

The best improvement over the Tanimoto alone is 170.6%; the lowest is 0.0%. The latter value is expected as this would reflect a weighting of 1.0 for the Tanimoto and 0.0 for the other coefficients. The best improvement rate over TSS is 151.4%, and the worst is −23% for a D2 antagonist. Averaged over all 55 searches, the improvement rate over the Tanimoto alone is 26.9%, and the improvement over TSS is 14.8%. Of the 55 searches carried out, 49 show an improvement over the Tanimoto alone, and 40 show an improvement over TSS.

Testing was carried out using the other half of the MDDR database as the data set and using four actives from each class with similar sizes to those used for training. The best weights resulting from the training stage were used for each combination of class and size range, and the measure of performance was, again, the number of actives retrieved in the top 5% of the ranked data set. In many cases, training resulted in several combinations for the best weighting. In such cases, training was carried out using one combination selected as being a good representative of those identified. Results for the same five classes are shown in Table 8.

Although, in a few cases, some impressive improvement rates are seen, the most impressive being a 1025% improvement for a single HIV-1 protease inhibitor over the Tanimoto alone and 462% over the TSS, this is offset by some poor results, the worst being −26% over the Tanimoto and −36% over the TSS. Overall, there is a net improvement of 49% over the Tanimoto and 26% over TSS, but these figures are

Table 9. Class-Based Training Improvement Rates over Tanimoto and TSS

	over the Tanimoto			over TSS		
class	TMod	TMAX	TSUM	TMod	TMAX	TSUM
renin inhibitor	6.7	8.7	0.7	5.5	7.4	0.2
angiotensin II AT1 ant.	42.6	37.2	29.1	5.7	3.6	3.5
thrombin inhibitor	33.1	72.7	45.2	29.3	67.4	43.1
substance P antagonist	5.8	29.3	8.8	4.5	22.4	8.1
5HT3 antagonist	26.6	38.0	29.9	5.2	16.7	11.4
5HT1A agonist	77.7	62.4	58.9	27.3	15.7	48.8
5HT reuptake inhibitor	26.1	18.1	29.8	22.0	14.4	27.8
HIV-1 protease inhibitor	22.8	33.3	13.5	21.4	31.7	9.9
D2 antagonist	25.6	28.6	43.7	−2.1	13.5	15.6
cyclooxygenase inhibitor	22.5	28.3	17.2	22.1	27.9	17.2
protein kinase C inhib.	22.2	11.9	14.9	22.2	11.9	7.6
overall average	**26.9**	**32.9**	**26.4**	**14.8**	**21.2**	**17.6**

Table 10. Summary of Improvement Rates for All Training Methods

	TMod	TMAX	TSUM
percentage of performances better than Tan	89.1	100	96.4
percentage of performances better than TSS	72.7	94.5	90.9
best individual improvement over Tan	170.6	283.8	202.5
least individual improvement over Tan	0.0	0.8	0.0
best individual improvement over TSS	151.4	283.8	202.5
least individual improvement over TSS	−23.0	−6.6	−0.5
overall average improvement over Tan	26.9	32.9	26.4
overall average improvement over TSS	14.8	21.2	17.6

Table 11. Class-Based Testing Improvement Rates over Tanimoto and TSS

	over the Tanimoto			over TSS		
class	TMod	TMAX	TSUM	TMod	TMAX	TSUM
renin inhibitor	7.7	7.6	9.3	5.4	4.5	4.4
angiotensin II AT1 ant.	33.6	39.1	40.4	17.2	1.4	5.8
thrombin inhibitor	111.6	42.4	113.2	111.4	42.4	76.3
substance P antagonist	5.2	3.3	4.0	5.15	3.3	4.0
5HT3 antagonist	35.4	15.7	22.8	−24.6	0.0	8.8
5HT1A agonist	61.9	33.6	48.3	9.1	5.2	22.0
5HT reuptake inhibitor	12.4	−0.3	0.8	9.7	−17.2	−1.1
HIV-1 protease inhibitor	270.9	80.2	13.9	130.3	39.8	4.4
D2 antagonist	0.0	40.1	29.4	−13.8	8.4	7.1
cyclooxygenase inhibitor	−2.4	23.0	7.7	−9.3	1.5	6.2
protein kinase C inhib.	4.7	7.8	13.9	−5.3	−3.3	3.4
overall average	**49.2**	**24.5**	**27.6**	**26.1**	**7.5**	**12.8**

Table 12. Summary of Improvement Rates for All Testing Methods

	TMod	TMAX	TSUM
percentage of performances better than Tan	63.6	88.6	90.9
percentage of performances better than TSS	52.3	72.7	84.1
best individual improvement over Tan	1025.0	250.0	42.9
least individual improvement over Tan	−26.0	−14.9	−43.2
best individual improvement over TSS	462.5	133.0	242.9
least individual improvement over TSS	−35.9	−39.1	−43.2
overall average improvement over Tan	49.2	24.5	27.6
overall average improvement over TSS	26.1	7.5	12.8

offset by a few high values. Of the 44 searches carried out, 28 show an improvement over the Tanimoto alone and 23 show an improvement over TSS.

TURBOMAX

TurboMAX training was carried out as for TurboModal, but steps of 0.25 were used to combine the weights. This is because TurboMAX takes much longer to train due to the increased number of searches involved. The same queries were used in the training stage as were used for TurboModal. Table 9 summarizes the training results for all three methodologies in terms of their improvement rate for each class over the Tanimoto alone and TSS; Table 10 summarizes their overall performance. TurboMAX clearly shows the best overall performance, with average increases of 32.9% over the Tanimoto and 21.2% over TSS. Indeed, all training results showed an improvement over the Tanimoto, with only three of the 55 searches being worse than TSS.

The same 44 queries from the TurboModal test stage were used against the other half of the MDDR. Table 11 summarizes the testing results for all three methodologies in terms of their improvement rate for each class over the Tanimoto alone and TSS; Table 12 summarizes their overall performance. Again, the combination of weights which best represented those identified for each class and size range in the training stage was used for each respective search. The results, although less impressive overall, show that this is a more robust methodology, with 39 out of the 44 searches producing better results than the Tanimoto alone, and 32 out of 44 showing improvement over TSS. Although the results for some of the classes are not as impressive as the TurboModal search, they are more consistent across the various activity types.

TURBOSUM

Since TurboSUM training takes a considerable amount of time, we only used 20,000 compounds as our training set and used incremental steps of 0.25 for the weights. We used the same 55 queries to train the weights as in previous training experiments. The results are not as impressive as those for TurboMAX, with slightly lower average improvement rates. They are, however, more consistent than TurboModal across all active types.

Again, the same 44 were used for testing the TurboSUM methodology. The improvement rates over Tanimoto and TSS are consistently high, more so than TurboMAX, and, although not as impressive for a few active classes in which TurboModal is superior, TurboSUM suffers minimal detrimental effect. Indeed, TurboSUM outperforms the Tanimoto in 40 of the 44 searches and outperforms TSS in 37 of the 44 searches.

Table 13. Comparison of Average Recall Performance with Bayesian Classifier Showing, for Each Class, the Number of Actives Retrieved in the Top 5% of the Test Set

	50K training set			20K training set	
class	TMod	TMAX	Bayesian classifier	TSUM	Bayesian classifier
renin inhibitor	517.75	517.50	523	526.00	519
angiotensin II AT1 ant.	341.25	352.25	411	355.25	385
thrombin inhibitor	185.75	145.00	294	183.75	228
substance P antagonist	172.75	169.00	450	169.00	294
5HT3 antagonist	67.00	72.25	292	74.25	284
5HT1A agonist	154.25	135.25	308	144.75	298
5HT reuptake inhibitor	40.00	36.25	104	35.00	69
HIV-1 protease inhibitor	104.25	102.25	256	88.25	178
D2 antagonist	37.25	52.00	140	47.50	131
cyclooxygenase inhibitor	46.74	47.50	205	51.25	194
protein kinase C inhib.	50.00	50.25	120	53.50	110

Table 14. Summary of Time Consumption for Training and Testing

method	training set size	step increment	training	testing
TMod	50,000	0.1	5 min 15.6 s	0 min 38 s
TMAX	50,000	0.25	76 min 13.1 s	0 min 5.4 s
TSUM	20,000	0.25	112 min 33.0 s	19 min 43.2 s

COMPARISON WITH STANDARD MACHINE LEARNING TECHNIQUES

In order to compare our methodology with standard machine learning techniques, we used the Bayesian classification model available in Scitegic's Pipeline Pilot[27] package. The same two sets of training data, 50,000 compound set and 20,000 compound set, were used to build models for all eleven classes in which compounds were characterized using BCI fingerprints. In the testing stage, the models were used to classify the test set (50,0000 compounds), and the number of actives in the top 5% of the classified test set was used to indicate recall performance. Table 13 compares the results for all eleven classes with the average recall for tests using weighted fusion.

Clearly, standard machine learning techniques are superior and outperform weighted fusion in all but one category here. However, the results presented here are for comparative purposes only, since the mode of operation of the two techniques is quite different. In weighted fusion, once the weights have been deduced for a particular query size, they can be applied to many similarity searches using reference structures of comparable size. In addition, subsequent search results can be used to further enhance the weighting schemes.

TIME CONSUMPTION FOR TRAINING AND TESTING

Table 14 shows the training and testing times for a single active compound, using a 1.8 GHz Pentium PC running Linux. Clearly, the TurboModal is a much faster operation, with training times of just over five minutes when using a 50,000 compound training set. This would be further improved with a reduction in step size. However, the more consistent methodologies are TurboMAX and TurboSUM, the latter being the most computationally expensive.

DISCUSSION

From the training experiments, a noticeable feature is that the best weights appear to reflect the distance between size distribution of the respective active class and that of the whole database. Figure 3 shows the size distributions in terms of number of bits set, standardized for comparison, of four active classes and the MDDR database. Clearly, the renin inhibitors and thrombin inhibitors are of a generally larger size than the bulk of the database and the 5HT1A agonists and 5HT reuptake inhibitors are generally smaller.

Table 15. Example Training Weights for Large and Small Active Classes

class	fusion method	For	SM	Tan	Rus
renin	TMod	0.0	0.0	0.4	0.6
	TMAX	0.0	0.0	0.5	0.5
thrombin	TMod	0.0	0.0	0.4	0.6
	TSUM	0.0	0.0	0.25	0.75
5HT1A agonist	TMod	0.5	0.0	0.5	0.0
	TMAX	0.25	0.25	0.5	0.0
5HT reuptake inhib.	TMAX	0.25	0.75	0.0	0.0
	TSUM	1.0	0.0	0.0	0.0

The best weights identified during training, as shown in Table 15, reflect this disparity. The larger classes, as expected, show a clear preference for those coefficients known to retrieve larger and medium sized compounds (i.e., Russell/Rao and Tanimoto), while the smaller classes show an affinity for those which are known to retrieve slightly smaller and much smaller compounds (i.e., Simple Match and Forbes), as seen in Figure 1.

The change of weights in Table 15 is quite striking and supports the notion that these coefficients can complement each other in terms of their ability to retrieve compounds across varying size ranges and that the correct choice of weighting these coefficients is an important factor for aiding retrieval.

The training methodology described above uses one example for each combination of active class and size range. In practice, there may be more than one such compound available, and the addition of further training results would enhance the weighting scheme accordingly. This may be applied in one of two ways. Either conduct the training methodology using the available actives individually and produce a consensus weighting based on these or combine the actives during the first iteration of the search in a multiple target approach similar to that described by Hert et al.[23]

The training stage often produces several combinations of weights, with several combinations producing the same level of retrieval. Our approach here has been to select an example which best represents those combinations identified

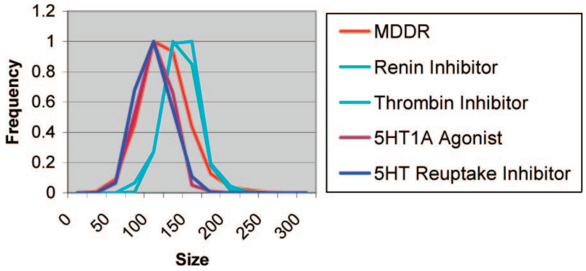


Figure 3. Standardized size distribution of classes within the MDDR database.

in training. An alternative approach might be to combine all best combinations using some consensus method and use this combined weighting in the testing stage.

CONCLUSIONS

A method has been described which attempts to automatically deduce the optimum set of weights which should be applied when combining four complementary similarity coefficients in order to improve the retrieval performance in searches for a given active class. These preliminary investigations have shown that there is considerable variation in the sets of compounds that are being retrieved by different similarity coefficients when used either individually or in combination using a data fusion methodology. The results show that a significant level of improvement can be achieved by training the weights of different coefficients and that training can be based on a single active compound of a given class and selected size range.

The studies strongly support the relationship between the choice of coefficient to use in a similarity search and the size distribution of the compounds retrieved. The study has identified four unique coefficients which clearly perform differently across different size ranges. The Russell/Rao is the more appropriate coefficient to use when the query is large, and the Forbes when the query is small. Moreover, it is possible to direct a search using alternative coefficients, or combination of coefficients, in order that relatively larger or smaller actives are retrieved. This may be during optimization stages when a lead may require an alteration in its bulk, for instance, in order to refine its ADME properties. Figure 1, for instance, shows that, almost independent of the size of the query itself, retrieval of like-sized actives can be successfully performed using the Tanimoto coefficient, whereas larger actives are more likely to be retrieved using the Russell/Rao, and smaller actives by the Simple Match or Forbes.

The choice of methodology is a trade off between time and robustness. Although Table 11 would appear to indicate generally high improvement rates for TurboModal, these are very much class-dependent with a few examples of inferior behavior. TurboSUM would appear to be the most robust methodology, but this is also the slowest for both training and testing. Training times can be reduced with the use of a smaller training set. Recent experiments using training sets of 10,000 and even 5000 structures for the TurboSUM methodology have produced comparable weightings with proportionally shorter training times. TurboModal is clearly the faster methodology, but it is also less consistent across all active classes. In particular, for the cyclooxygenase inhibitors and, to a lesser degree, 5HT3 antagonists and 5HT reuptake inhibitors weightings identified during the training stage varied considerably even within the same size range, the effect of which is a considerable difference in performance in the testing stage, as illustrated in Table 8. TurboSUM training produces more consistent weightings across all classes and therefore improved consistency for searching.

ACKNOWLEDGMENT

We thank Daylight Chemical Information Systems Inc. for funding for this project and for the provision of fingerprinting software and MDL for the provision of the MDDR database. We also thank Tripos Inc., Accelrys Software Inc., and Digital Chemistry for the provision of further software used during this project.

REFERENCES AND NOTES

(1) *Virtual Screening in Drug Discovery*; Alvarez, J., Shoichet, B., Eds.; CRC Press: Boca Raton, FL, 2005.

(2) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(3) Willett, P. Similarity methods in chemoinformatics, 2008. *Ann. Rev. Inf. Sci. Tech.* **2009**, *43*, 3–71.

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

(5) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Diversity* **2006**, *10*, 39–79.

(6) Weininger, D.; Delany, J. J.; Bradshaw, J. A Brief History of Screening Large Databases. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html#RTFToC77 (accessed July 29, 2008).

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

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

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

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

(11) 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.

(12) 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.

(13) Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327–352.

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

(15) *Mathematical techniques in multisensor data fusion*; Hall, D. L., Ed.; Artech House: Northwood, MA, 1992.

(16) *Sensor and data fusion concepts and applications*, 2nd ed.; Klein, L. A., Ed.; SPIE Optical Engineering Press: Bellingham, 1999.

(17) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887–2900.

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

(19) Flower, D. R. On the properties of bit string based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1988**, *38*, 379–386.

(20) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163–166.

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

(22) Daylight Chemical Information Systems, Inc. http://www.daylight.com (accessed Nov 10, 2008).

(23) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based

virtual screening using nearest-neighbor information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(24) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 462–470.

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

(26) Digital Chemistry. http://www.digitalchemistry.co.uk (accessed Nov 10, 2008).

(27) Accelrys Software Inc. Scitegic Platform. http://accelrys.com/products/scitegic (accessed Nov 10, 2008).

CI800292D