Conditional Probability: A New Fusion Method for Merging Disparate Virtual Screening Results

John W. Raymond,* Mehran Jalaie, and Mary P. Bradley

Pfizer Global Research and Development, Discovery Technologies, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, Michigan 48105

Received October 27, 2003

This paper introduces a new consensus scoring approach for merging the results of different virtual screening methods based on conditional probabilities. The technique is experimentally evaluated using several ligand-based virtual screening methods and compared to two variations of the established Sum-rank fusion method where it performs as well or better than the Sum-rank methods. Our experiments confirm that consensus scoring increases the number of active compounds retrieved with respect to the best individual methods on average.

INTRODUCTION

The use of similarity-based methods in pharmaceutical virtual screening (VS) has been well-documented[1−3] and come in a wide variety of techniques. The lead-compound-based methods typically compare a query compound possessing a known or desired chemical property to a list of other compounds in a database and assigning a numerical value depending upon the perceived level of similarity. The selected list of database compounds are then sorted in order of nonincreasing value of similarity with the highest ranked compounds presumed to boast a higher probability of possessing a chemical property similar to that of the query—the much referenced "similarity principle".

Since it is difficult to establish which method may perform the best given a particular instance, data fusion methods, also referred to as consensus scoring, have been recently employed for merging virtual screening results from different similarity methods.[4,5] The potential benefits of a successful data fusion approach are 2-fold. First, it helps reduce the uncertainty involved in selecting the appropriate VS method, analogous to stock market indexing. Given a set of plausible VS results from various methods for a particular query, it is most likely that a user will not know a priori which method's results are most appropriate and could just as easily select an inferior search result as select a successful one. Second, useful data fusion approaches can potentially result in a merged result that is superior to any of the individual input method results and thus is able to extract useful information from all input lists, including inferior methods.[6]

Of the published fusion methods, the Sum-rank method has been shown to be among the most successful.[5,7−9] In this paper, we introduce a new VS hit-list fusion method based on conditional probabilities. It is compared directly to the established Sum-rank method. The individual VS methods chosen for the study are Daylight[10] and BCI[11] fingerprints, 2D MCS[7,12] and 3D MCS,[13] Tripos' dbtop,[14] and Open Eye's ROCS.[15] The following analysis seeks to determine which combination of the VS method produces the best results as well as to establish whether the proposed fusion method provides any benefit over the established Sum-rank method.

* Corresponding author phone: (734)622-3015; fax: (734)622-2782; e-mail: john.raymond@pfizer.com.

METHODS

Daylight Fingerprints. Daylight fingerprints[10] were used for 2D bit-string similarity comparisons where the bits represent 2D graph path lengths between atoms. The default specifying 2048 bits was used. Similarity was calculated using the Tanimoto coefficient

$$S_{12} = \frac{c}{a + b - c} \tag{1}$$

where a and b are the number of bits set in the first fingerprint and second fingerprint being compared, respectively, and c is the number of bits that are set in both fingerprints.

BCI Fingerprints. BCI fingerprints[11] were used to represent fragment-based 2D bit-string similarity. The default 1092 bit fragment dictionary was used, and similarity was calculated using the Tanimoto coefficient.

2D MCS. This method is a 2D graph-based similarity measure which uses the maximum common substructure (MCS) as a measure of similarity between two chemical structures represented as graphs. The measure which ranges from 0 to 1 is a modified version of the Kulczynski similarity coefficient

$$S_{12} = \frac{|G_{12}|\cdot(|G_1| + |G_2|)}{2|G_1|\cdot|G_2|} \tag{2}$$

where $|G_{12}|$ is the MCS between the two structures, $|G_1|$ and $|G_2|$, being compared. The values of $|G_{12}|$, $|G_1|$, and $|G_2|$ are calculated as specified in the literature.[7]

3DMCS. The 3D MCS is similar to the 2D MCS approach and uses the same underlying algorithm,[12] but instead of defining similarity using the 2D MCS, it uses the 3D MCS.[13] The 3D MCS is the largest subset of geometrically consistent

10.1021/ci034234o CCC: $27.50
Published on Web 01/21/2004

atoms common to both structures being compared. The similarity measure employed is a modified version of the Tanimoto coefficient

$$S_{12} = \frac{|G_{12}|}{2\cdot|G_1| + |G_1 - G_2| - |G_{12}|} \quad (3)$$

dbtop. dbtop[14] is a shape-based similarity application developed by Tripos[16] that fragments the molecules being compared into smaller fragments. The shapes of these fragments are called topomers. The steric similarity of two structures is described as the squared sum of differences in the steric field values of their constituent topomers.[14] dbtop returns a distance value indicating the level of similarity between two structures ranging from 0 (a perfect match) to an arbitrarily large number. The application was run using default parameters unless otherwise specified.

ROCS. ROCS is software designed by OpenEye[15] for 3D database searching. It relies on rapid Gaussian shape overlay of compounds. Specifically, molecules are aligned in such a way to maximize their Gaussian overlap volume, and a shape-based Tanimoto score is assigned to measure the shape similarity between a probe and any molecule in the database (the concept of shape and volume overlaps in the Gaussian-based context are related).[17,18] Of the methods tested, only ROCS is conformationally dependent. The application was run using default parameters unless otherwise specified.

Sum-Rank Method. The Sum-rank method is a simple rank-based fusion technique. It is described by eq 4

$$SUM_x = \sum_{i=1}^{N} R_i(x) \quad (4)$$

where $R_i(x)$ indicates the rank position of compound x in the ranked result (i.e., hit-list sorted in nonincreasing order) from VS method i and N is the number of VS methods under consideration. The calculated values of SUM_x for each compound in the searched database are then sorted in nonincreasing order. The top ranked compounds are then considered as the fused hit-list.

Since there is typically no discernible correlation between the individual similarity values for the various VS methods, it is not feasible to directly merge the various VS results using the underlying similarity measures.[5] The Sum-rank method avoids this problem by using the ordered rankings.

However, the use of rank positions introduced a potential limiting feature. Rank collisions occur when multiple compounds in any of the input VS method hit-lists have the same similarity score. This creates a problem because compounds with identical similarity scores will have different rank positions within the sorted list. We consider the case of multiple compounds with the same similarity score by implementing the Sum-rank method using two different decision criteria. In the first, the rank positions of compounds with identical similarity scores are randomly assigned by the sorting algorithm, and in the second, the rank positions are established by assigning compounds with equal similarity scores the same rank position. For instance, the ranking for the set of scores (0.8, 0.7, 0.7, and 0.4) would be (1, 2, 3, 4) and (1, 2, 2, 3) in the random and equivalent ranking circumstances, respectively.

Conditional Probability (CP) Method. The proposed fusion algorithm approaches the consensus scoring problem by weighting each compound in a VS run with an activity-based conditional probability, $P(A|S_x)$. $P(A|S_x)$ is the probability that a database compound (x) is active (A) given a similarity score (S_x) with respect to an active query compound. Assuming that each compound in a VS hit-list has been assigned a value of $P(A|S_x)$, then a straightforward fused ranking can be established by assigning a comparative score to each compound screened using the relation in eq 5

$$CP_x = \prod_{i=1}^{N} P(A|S_x^i) \quad (5)$$

where S_x^i is the value of S_x for the VS method i and N is the number of screening lists being fused.

The calculated CP_x scores are then sorted in nonincreasing order with the top ranked compounds assumed to have the highest probability of being active. In a probabilistic interpretation, the CP_x score assumes that the individual activity probabilities represent independent events. This is obviously not the case since each similarity score describes the same pair of compounds (i.e., outcome). A more appropriate ranking score would be $CP_x = P(A|S_x^1,...,S_x^N)$. However, considering joint probabilities significantly increases the difficulty of empirically obtaining estimates of the necessary probability distributions.

While the assumption of independent events has the potential to be a substantial limitation, it is mitigated to some extent by the fused score ordering process. Consider for instance the hypothetical situation where one wants to know which of two jewel thieves has the highest likelihood of being arrested given two observed events. One jewel thief has a given probability of being arrested if he leaves a fingerprint at the crime scene, $P(A|S^1) = 0.5$ and a probability of $P(A|S^2) = 0.3$ that he will be arrested if he leaves any DNA at the crime scene because he was previously convicted. The other thief has probabilities of arrest, $P(A|S^1) = 0.2$, and $P(A|S^2) = 0.2$, respectively. Clearly, the actual probabilities that each thief will be arrested if he leaves both a fingerprint and DNA at the crime scene is not $0.5 \times 0.3 = 0.15$ and $0.2 \times 0.2 = 0.04$, respectively. However, it is plausible to believe that the two calculated values can be used to establish a relative ranking of which scenario is most likely to result in an arrest (i.e., the first thief is most likely to be arrested).

Another interesting feature of this approach is that it can be used independently of the "similarity principle". In the CP method, the lists are sorted in order of increasing probability of being active rather than similarity. In contrast to the Sum-rank method, this allows the proposed method to merge disparate data even in the situation where the probability of being active is not monotonically increasing (decreasing) with increasing (decreasing) values of an observed score/measurement.

To perform the proposed conditional probability (CP) fusion, it is necessary to obtain the values of $P(A|S_x)$ for the compounds being screened. This is estimated with the aid of Bayes' rule. Using Bayes' rule, we can define $P(A|S_x)$ as

$$P(A|S_x) = \frac{P(S_x|A)\cdot P(A)}{P(S_x)} \quad (6)$$

where $P(S_x)$ is the unconditional probability that a similarity score of S_x will occur, and $P(A)$ is the unconditional probability that the comparison involves a database compound exhibiting similar activity to the query compound. $P(S_x|A)$ is the conditional probability that a similarity value of S_x will occur given that the comparison involves both a database and query compound of like activity.

We have recently discovered that Manmatha et al.[19] have also employed Bayes' rule as a weighting mechanism in their mixture model procedure for fusing text document retrieval results. However, the proposed method is significantly simpler than their approach as it does not require the estimation of mixing parameters or multiple distributions for each similarity search method.

Prior to comparing the CP and Sum-rank methods, it is first necessary to establish a correlation for $P(A|S_x)$ given a calculated similarity score S_x. Fortunately, the compulsory parameters (as identified by Bayes' rule) are readily calculable using a frequency approach given a training set of sufficient size and quality. The unconditional probability, $P(S_x)$, of encountering a similarity value of S_x for a given VS method can be determined by performing a large number of random pairwise comparisons and the fitting a probability distribution to resulting values. The conditional probability distribution $P(S_x|A)$ can also be empirically estimated by first clustering a large number of compounds into groups of like activity and then performing all intracluster, pairwise similarity comparisons.

Since we are primarily concerned with relative rankings, it is assumed that the suggested mechanisms for estimating $P(S_x)$ and $P(S_x|A)$ are reasonably independent of the selected training set in the sense that how the ratio $P(S_x|A)/P(S_x)$ for one VS method varies with respect the ratio $P(S_x|A)/P(S_x)$ for another VS method is consistent across other sufficiently large data sets. However, using the training set approach to estimate the unconditional probability of activity, $P(A)$, is much less plausible. Fortunately, it is not necessary to determine an accurate value for $P(A)$. Since $P(A)$ is independent of any particular VS scheme and CP is a ranking method not dependent upon the absolute values of $P(A|S_x)$, we can assume $P(A)$ to be an arbitrary constant as it will be a constant coefficient in all values of CP_x regardless of the VS method. Therefore the conditional probability, $P(A|S_x)$, used in CP_x can be condensed to

$$P(A|S_x) \cong \frac{P(S_x|A)}{P(S_x)} \tag{7}$$

As justification, substituting eq 6 into eq 5 yields

$$CP_x = P(A)^N \cdot \prod_{i=1}^{N} \frac{P(S_x^i|A)}{P(S_x^i)} \tag{8}$$

Since $P(A)$ is assumed constant across all VS methods and concerned only with relative rankings, this equation further simplifies to

$$CP_x = \prod_{i=1}^{N} \frac{P(S_x^i|A)}{P(S_x^i)} \tag{9}$$

Although the calculated CP score is intended primarily to establish relative rankings, a qualitative interpretation of the score can be obtained by taking the Nth root of the score. This can be loosely interpreted as the geometric mean of individual VS method probability enrichment values. In other words, $N\sqrt{CP_x}$ may provide a crude consensus enrichment factor over random selection for the VS methods used; however, this value is not used directly in the experiments performed here.

CP Fusion Calibration. Before validation simulations can be performed comparing the CP and Sum-rank methods, correlations for the necessary probability distributions $P(S_x)$ and $P(S_x|A)$ must be calculated. For this procedure, we use a filtered subset of the MDDR database consisting of 81 796 nonduplicated compound structures and their associated activities. The filtered subset consists of those compounds with a molecular weight greater than 100 and less than 600 with 15 or fewer rotatable bonds.

For all methods except ROCS, the procedure used to calculate $P(S_x)$ consisted of randomly selecting a subset of 10 000 compounds from the MDDR training set and then performing all pairwise comparisons. The 10 000 compound subset resulted in approximately 50 million pairwise comparisons per VS method for BCI, Daylight, 2D MCS, and 3D MCS. Since the dbtop distance calculation is asymmetric (i.e., A→B ≠ B→A), the pairwise calculation procedure for dbtop resulted in approximately 94 million comparisons (less than 100 million since dbtop was not able to make some comparisons). The computed similarity values from each of these methods were recorded as frequency counts in predefined similarity range bins. For all methods except dbtop, the bins were labeled with similarity ranges in increments of 0.01. For instance, the first two bins were 1.0−0.99 and 0.99−0.98, respectively. dbtop distances were binned in increments of 10 (i.e, 1−10, 10−20, ...). Bin sizes were selected so as to allow as high resolution as possible while also ensuring that each bin sufficiently populated. Whenever a similarity value fell within the range of a specified bin, the bin frequency counter was incremented by one. These frequencies were then converted to probabilities by dividing by the total number of pairwise comparisons performed.

Due to time constraints for the case of ROCS, we randomly selected 3200 compounds from the 81 796 compound data set and then performed all pairwise comparisons. Each compound in the data set was represented by OMEGA[15] generated conformations using *-maxconfs* 400 (maximum number of conformers). Each comparison, therefore, consisted of a Concord[16] generated query mol2 structure compared to multiple, OMEGA conformer poses. A Concord generated query structure was used since a bound crystal structure conformation was not available and efficiency considerations precluded using multiple query conformations. The resulting 5.1 million comparisons (not counting comparisons involving multiple conformers in the database file as separate comparisons) were then binned in 0.01 increments as previously described.

To calculate $P(S_x|A)$, each compound was organized into groups of similar activity (i.e., same MDDR activity code). Compounds with multiple activity codes were assigned to multiple activity groups. Only activity groups containing 250 or fewer compounds were considered. A decision of a 250

Table 1. Optimized CP Scoring Function Parameters for Each Virtual Screening Method[a]

	Daylight	BCI	2D MCS	3D MCS	dbtop	ROCS
a	−9.223	−13.46	−213.2	−15.16	1175	0.00
b	0.0220	0.0248	202.4	0.0284	40.51	1.00
c	11.01	11.19	25.20	11.34	−116.6	32.70
d	0.2642	0.3253	3.930	0.2484		1.00
e	0.2391	0.1994	0.4824	0.3		0.7678

[a] Daylight, BCI, 2D MCS, 3D MCS, and Rocs use eq 10. dbtop uses eq 11.

member cutoff was based on our judgment of the composition of the MDDR database. Since there is no assurance that compounds possessing the same activity code bind in the same mode or even to the same protein target, we decided that activity groupings containing more than 250 members carried an unnecessary risk of grouping structurally unrelated compounds. For each activity group with 250 or fewer members, all intragroup pairwise similarities were calculated and binned as previously described. This resulted in approximately 1.5 million pairwise comparisons for the 474 activity groups with 250 or fewer members. This approach was followed for each VS method under consideration.

Having established discrete probability distributions for $P(S_x)$ and $P(S_x|A)$ for each screening method, the CP scoring function, an approximation of $P(A|S_x)$, was calculated for each binned similarity range increment by dividing $P(S_x|A)$ by $P(S_x)$. To make the CP scoring function more convenient to use, curves were fitted to the binned values using Sigma Plot 7.0. A five parameter exponential curve was found to be sufficiently general to provide a good fit for all of the methods considered with the exception of dbtop. The equation for the CP scoring function for BCI, Daylight, 2D MCS, 3D MCS, and ROCS is

$$P(A|S_x) = a + b \cdot e^{c \cdot (x-e)^d} \tag{10}$$

The best curve fit for the dbtop distribution was found to be a three parameter Gompertz equation

$$P(A|S_x) = a \cdot e^{-e^{-(x-b)/c}} \tag{11}$$

The optimized parameters for each method's scoring function are listed in Table 1. Figure 1 depicts the fitted curves for each $P(A|S_x)$ scoring function.

VALIDATION

MDDR Validation. A set of 100 query structures was selected at random from the data set of 81 796 structures subject to the constraint that each selected compound could be present in only one of the MDDR activity groups represented in the data set. This constraint was used to unbias the evaluation process. We wanted to avoid the situation where a compound with multiple associated activities was used as a query, and one screening method was biased toward one of the activity groups while another method preferentially selected compounds from another equally valid activity group. A total of 51 distinct activity groups were represented by the 100 query structures.

Table 2. Results of the Top Ranked Compounds for Each Virtual Screening Method[a]

	Daylight	BCI	2D MCS	3D MCS	dbtop	ROCS	Max(*all*)
$\bar{a}_{150}$	35.8	35.8	35.1	38.5	32.0	12.9	48.4
ΔA_{150}	29.3	29.3	30.5	26.9	40.0	69.1	
$\sigma_{\Delta A_{150}}$	27.4	28.4	23.9	28.3	34.4	29.2	

[a] $\bar{a}_{150}$ is the average of the number of active compounds retrieved in the top 150 compounds. $\Delta A_{150}(\sigma_{\Delta A_{150}})$ is the average disparity (standard deviation of ΔA_{150}) between each individual method and the best overall method for each query.

For validation purposes, the data set of 81 796 structures was searched using each VS method with the 100 query structures. The method's ability to retrieve similarly active structures was evaluated using the number of active compounds retrieved in the 150 top ranked compounds. The relatively small hit-list size of 150 compounds represents approximately 0.2% of the database and was selected so that quantity of candidate structures resulting from the screening was manageable enough to potentially permit IC^{50} confirmation testing on a scaled-up basis. For instance, 0.2% of a 2 million molecule database would require that 4000 compounds need experimental follow-up.

Table 2 lists the average number of active compounds retrieved per search for each method. For reference, the hypothetical best case scenario where the best result across all six methods for each query was selected is also presented. In addition, the relative discrepancy between each individual method and the best overall method for each query was calculated as

$$\Delta A_{150} = 100 \cdot \frac{a^x - \max_{i=1}^{k}\{a^i\}}{\max_{i=1}^{k}\{a^i\}} \tag{12}$$

where k is the number of single method rankings being considered, and a^x and a^i indicate the number of active compounds in the top 150 compounds for the rankings under consideration. The average and standard deviation of ΔA_{150} over all 100 queries for each method is also presented in Table 2.

The data in Table 2 indicate that there was no single dominant method across all 150 queries and that there is significant variability in which method performs best for an individual query. All of the methods tested performed comparably with the exception of ROCS. The poorer performance of ROCS is most likely due to the lack of a protein-bound ligand conformation rather than any deficiency in the method. These results bolster the notion that there is not an obviously superior screening method and that there is sufficient basis for a technique for effectively merging the virtual screening results to significantly enrich the recall of active compounds.

All 2- and 3-tuples were used to evaluate the proposed fusion procedure. This resulted in 35 fusion experiments (15 2-tuples and 25 3-tuples) for each query. For comparative, purposes every fusion experiment consisted of the Sum-rank method (random and equivalent cases) and the proposed CP fusion method. The Sum-rank and CP fused rankings were compared using the average number of active compounds retrieved over all 100 queries.

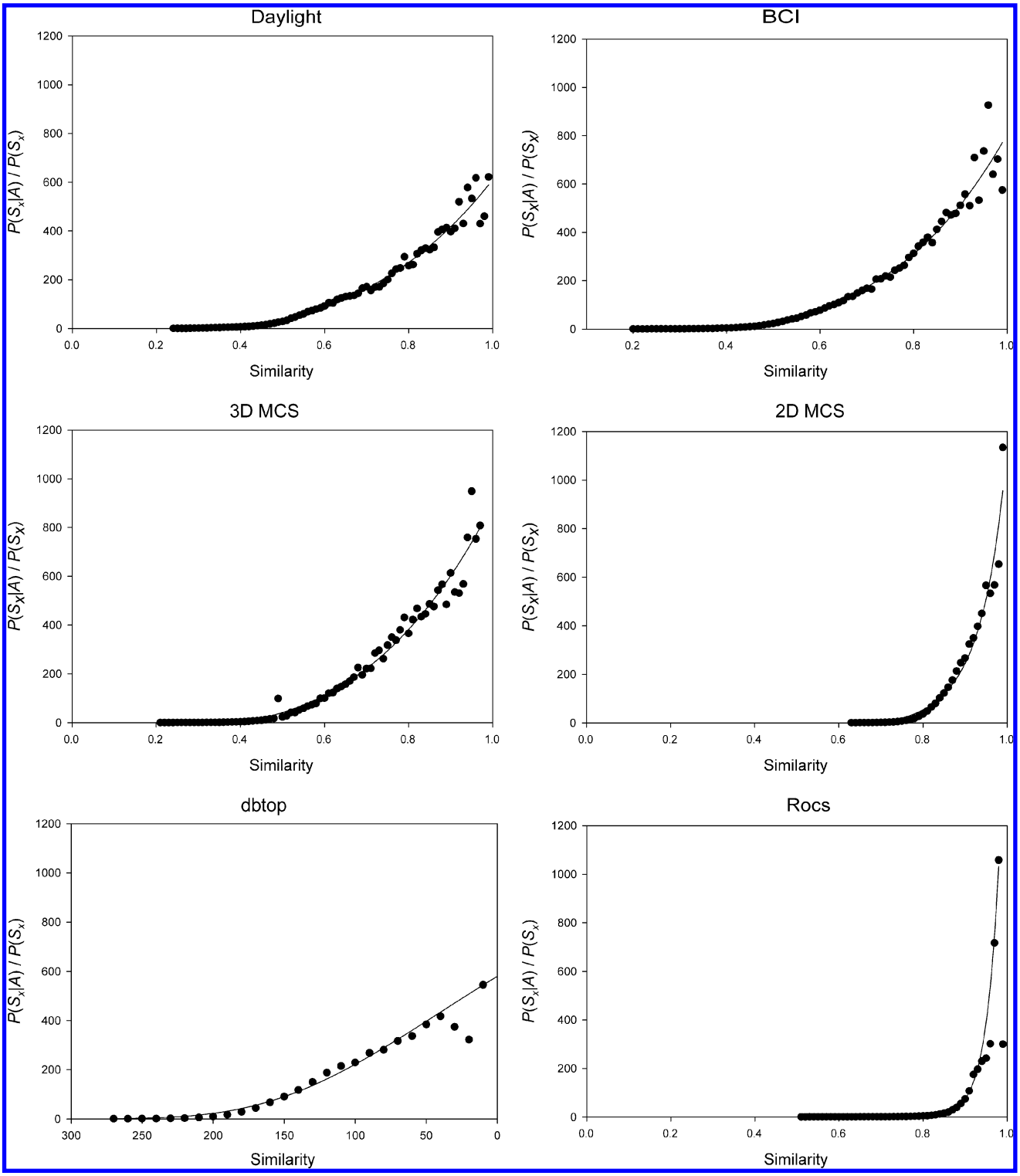

Figure 1. Curve fits for $P(S_x|A)/P(S_x)$ data for each screening method.

The first question to be addressed by the data in Tables 3 and 4 is whether there are any systematic differences between the fusion methods. This is accomplished using a paired sign test[20] between the mean recall values listed in Tables 3 and 4. Intuitively we assume that the equivalent assignment is superior to the random assignment Sum-rank method since it accounts for rank collisions. For this comparison the null hypothesis is that there is no difference between the two distributions, and the alternative hypothesis is that the equivalent assignment is superior to the random assignment. Using the one-tailed sign test, the null hypothesis can be rejected with a significance level of 0.0005 for both the 2-tuple and 3-tuple scenarios, clearly indicating the superiority of the equivalent assignment alternative. Repeating the sign test for the equivalent assignment Sum-rank and CP methods, where the alternative hypothesis is that the CP method is superior to the Sum-rank, indicates that the CP method is clearly superior to the Sum-rank method for the 3-tuple case with a significance level of 0.0002. However, for the 2-tuple case, the null hypothesis cannot be rejected, and it cannot be concluded based on the data for this case that the CP method is better than the Sum-rank method.

Table 3. Average Number of Active Structures Retrieved for All 2-Tuple Fusion Simulations Each of the Three Fusion Methods under Consideration

sum-rank (random assignment)		sum-rank (equivalent assignment)		CP-rank	
$\bar{a}_{150}$	methods	$\bar{a}_{150}$	methods	$\bar{a}_{150}$	methods
38.9	BCI/3D MCS	***39.5***	***BCI/3D MCS***	***40.9***	***Daylight/3D MCS***
38.8	2D MCS/3D MCS	***38.78***	***Daylight/3D MCS***	***40.8***	***BCI/3D MCS***
38.8	Daylight/3D MCS	***38.25***	***dbtop/3D MCS***	39.7	2D MCS/3D MCS
37.4	BCI/2D MCS	37.93	BCI/2D MCS	***38.8***	***BCI/2D MCS***
36.8	dbtop/3D MCS	37.92	2D MCS/ 3D MCS	38.7	dbtop/3D MCS
36.8	Daylight/2D MCS	37.85	Daylight/dbtop	38.1	Daylight/2D MCS
36.6	Daylight/dbtop	37.28	Daylight/ROCS	37.8	BCI/Daylight
36.6	BCI/Daylight	37.23	BCI/dbtop	37.8	3D MCS/ROCS
36.0	2D MCS/dbtop	***37.1***	***Daylight/2D MCS***	36.7	Daylight/ROCS
35.9	BCI/dbtop	37.05	BCI/Daylight	36.5	2D MCS dbtop
25.2	BCI/ROCS	36.65	BCI/ROCS	36.4	BCI/ROCS
25.1	Daylight/ROCS	36.63	2D MCS/dbtop	36.3	Daylight/dbtop
23.8	2D MCS/ROCS	35.2	2D MCS/ROCS	36.3	BCI/dbtop
20.1	dbtop/ROCS	28.78	3D MCS/ROCS	35.2	2D MCS/ROCS
18.5	3D MCS/ROCS	27.2	dbtop/ROCS	26.2	dbtop/ROCS

Table 4. Average Number of Active Structures Retrieved for All 3-Tuple Fusion Simulations Each of the Three Fusion Methods under Consideration

sum-rank (random assignment)		sum-rank (equivalent assignment)		CP-rank	
$\bar{a}_{150}$	methods	$\bar{a}_{150}$	methods	$\bar{a}_{150}$	methods
38.7	BCI/2D MCS/ 3D MCS	40.1	***BCI/ 2D MCS/3D MCS***	41.4	***Daylight/dbtop/3D MCS***
38.5	BCI/Daylight/ 3D MCS	39.7	***BCI/dbtop/3D MCS***	41.4	***Daylight/3D MCS/ROCS***
37.9	Daylight/2D MCS/ 3D MCS	39.4	***Daylight/dbtop/3D MCS***	41.1	***BCI/3D MCS/ROCS***
37.9	2D MCS/ dbtop/3D MCS	39.3	***BCI/3D MCS/ROCS***	41.1	***BCI/dbtop/3D MCS***
37.5	BCI/dbtop/ 3D MCS	39.3	***BCI/Daylight/3D MCS***	40.9	***BCI/2D MCS/3D MCS***
37.5	Daylight/dbtop/3D MCS	39.2	***Daylight/3D MCS/ROCS***	40.7	BCI/Daylight/3D MCS
37.4	BCI/Daylight/2D MCS	38.7	Daylight/2D MCS/3D MCS	40.6	***Daylight/2D MCS/3D MCS***
37.0	BCI/2D MCS/ dbtop	38.6	BCI/2D MCS/ROCS	40.4	2D MCS/3D MCS/ROCS
36.9	BCI/Daylight/dbtop	38.5	BCI/2DMCS/dbtop	40.2	2D MCS/dbtop/3D MCS
36.9	Daylight/2D MCS/ dbtop	38.3	Daylight/dbtop/ROCS	39.8	***BCI/Daylight/2D MCS***
29.8	BCI/2D MCS/ ROCS	38.2	2D MCS/dbtop/3D MCS	39.2	BCI/2D MCS/ROCS
29.0	BCI/Daylight/ROCS	38.1	BCI/Daylight/2D MCS	39.0	BCI/2D MCS/dbtop
28.5	BCI/dbtop/ROCS	37.9	2D MCS/3D MCS/ROCS	38.8	Daylight/2D MCS/dbtop
28.4	Daylight/2D MCS/ ROCS	37.9	Daylight/2D MCS/dbtop	38.7	Daylight/2D MCS/ROCS
28.1	Daylight/dbtop/ROCS	37.8	BCI/Daylight/dbtop	38.5	BCI/Daylight/ROCS
27.4	2D MCS/dbtop/ROCS	37.7	BCI/Daylight/ROCS	37.9	BCI/Daylight/dbtop
27.1	BCI/3D MCS/ ROCS	37.6	Daylight/2D MCS/ROCS	37.8	dbtop/3D MCS/ROCS
26.8	Daylight/3D MCS/ ROCS	37.3	BCI/dbtop/ROCS	37.5	Daylight/dbtop/ROCS
25.9	2D MCS/ 3D MCS/ ROCS	36.6	2D MCS/dbtop/ROCS	37.1	BCI//dbtop/ROCS
22.7	dbtop/3D MCS/ ROCS	33.5	dbtop/3D MCS/ROCS	36.9	2D MCS/dbtop/ROCS

Having established the relative effectiveness of the fusion methods, it remains to be considered whether there are any combinations of screening methods which are systematically superior across fusion methods or whether they are fusion method dependent. For this analysis, the Spearman rank correlation coefficient[20] was employed. The Spearman coefficient is used to test the null hypothesis of "no association" between two distributions. Therefore, it can be used to determine whether there is a positive correlation between the resultant rankings from both of the Sum-rank alternatives and the CP method. The Spearman rank correlation tests between the rankings in Tables 3 and 4 indicate that there is a strong correlation between the rankings, and it can be concluded that there exist screening method combinations that perform consistently well across fusion methods.

While the rankings in Tables 3 and 4 provide a rough gauge of the effectiveness of each screening combination relative to the other screening combinations, it cannot be dependably used to establish a "best" combination for each fusion method due to random behavior in the retrieval results. We prefer to make the assumption that based on the data generated there are a subset of combinations which cannot be excluded from consideration as being among the best. The subsets of best combinations were determined by first assuming that the top ranked combination was one of the best combinations for a particular ranking. Then a paired, Wilcoxon signed-rank test[20] was used to compare the top-ranked combination with all other combinations in its respective ranking. The Wilcoxon signed-rank test is used to determine whether a difference exists between two distributions. It not only considers the number of positive or negative sign differences between the pairs (as in the sign test), but it also considers the relative magnitude of the differences.

The paired, Wilcoxon signed-rank test (one tail) consisted of comparing the number of active compounds retrieved for each of the 100 queries from the top-ranked combination with all other combination in each ranking, respectively. The null hypothesis is that no difference in the distribution of active compounds retrieved exists, and the alternative hypothesis is that the top-ranked combination performed better. Focusing on just the equivalent assignment Sum-rank

and the CP cases since the random assignment was found to be inferior to the equivalent assignment for both the 2-tuple and 3-tuple alternatives, the italicized cells in Tables 3 and 4 indicate the screening combinations for which the null hypothesis could not be excluded using a significance level of 0.1. This does not necessarily indicate that all highlighted methods within an individual ranking are equivalent, but that this possibility cannot be excluded.

For 2-tuple fusions, it appears that the best combinations consist of the 3D MCS and one of the two fingerprint methods, BCI or Daylight, as both of these combinations are present in "best" subsets of both the Sum-rank and CP fusion methods. For the 3-tuple combinations, there were five combinations in common between the two fusion techniques. Four of the five "best" combinations consisted of the 3D MCS, a fingerprint method (BCI or Daylight), and a shape-based method (dbtop or ROCS). The other consisted of the BCI fingerprint, 2D MCS, and the 3D MCS. Any further refinement as to which of these method combinations are actually the best would require a larger sample involving more query simulations.

"Real-World" Validation. To simulate how well the proposed fusion method would perform as a "real-world" virtual screening tool, we assembled a query set of structures of known activity from the literature to search the 81 796 compound subset of the MDDR. Since our MDDR data set was last updated in 2002, we limited our query structures to novel structures published in 2003. By limiting the query structures to those published in the *Journal of Medicinal Chemistry* in 2003, we decrease the likelihood that the MDDR data set will contain related analogues, thus mimicking the situation where a researcher obtains a structure of interest from the journal or patent literature and wishes to search in-house or vendor databases which have a very low probability of containing similar analogues.

One or more compounds were selected from each of the published articles referencing at least one valid MDDR activity present in the 81 796 compound data set. Although each article typically presented numerous potential candidates, no systematic rule was applied when selecting each query compound. Note that not all MDDR activity codes are present in the 81 796 compound data set due to the previously described filtering process. Figure 2 depicts the 26 compounds selected and their associated MDDR activity code(s). These 26 query compounds were then used to search the 81 796 compound data set to determine how well the CP method was able to discover other compounds of like activity, thus operating at the low end of similarity effectiveness for the individual screening methods (see Figure 1). For this simulation, the 3-tuple combination of Daylight/dbtop/3D MCS was selected as it was among the best performing of those tested in the previous analyses.

Table 5 lists the number of active structures retrieved in the top 150 compounds in the 26 searches for each method. A 150 structure cutoff (approximately 0.2% of the database) was used in an attempt to ensure scalability for the observed results as previously described.

The number of active structures expected to be retrieved via random selection is also listed. These results indicate that the individual methods under consideration perform surprisingly well and that there is a marked disparity in the distribution of active structures retrieved relative to the each query—further highlighting the uncertainty associated with method selection for a given query. Table 5 also lists the results of the CP fusion for the three methods. The data clearly indicate that the fused lists are able to retrieve more actives relative to the individual methods for these query structures and just as importantly, also reducing the uncertainty associated with individual screening method selection.

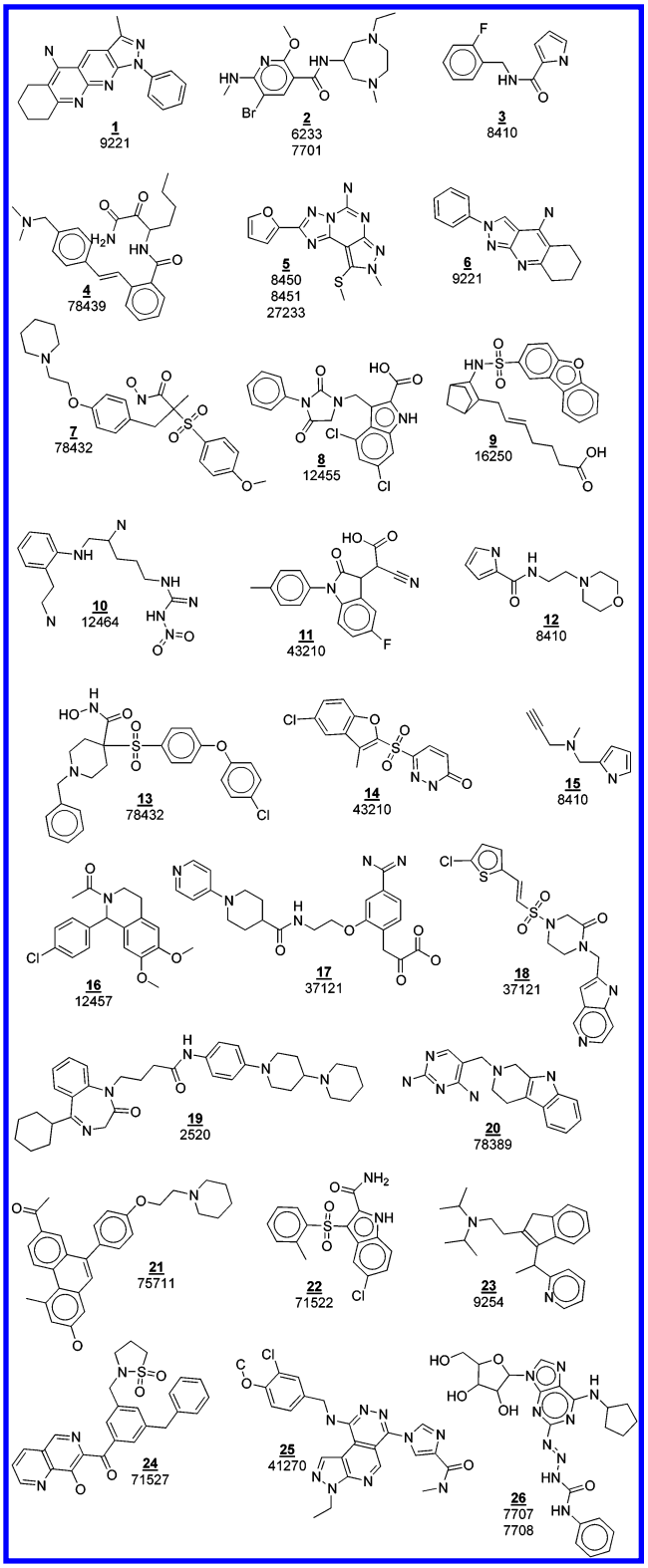

Figure 2. Chemical structures used in "real-world" screen. Each structure annotation includes the structure ID and its designated MDDR activity codes.

Table 5. Number of Active Structures Retrieved for the 26 "Real-World" Queries in the Top 150 Ranked Compounds

	number of active compounds retrieved				
compd	Daylight	dbtop	3D MCS	random	Daylight/ dbtop/3D MCS
1[21]	84	0	3	1.1	67
2[22]	18	17	57	2.1	48
3[23]	0	0	0	0.14	0
4[24]	3	0	0	0.066	2
5[25]	68	32	41	0.49	70
6[21]	88	0	10	1.1	58
7[26]	7	0	9	0.59	11
8[27]	47	69	56	2.3	71
9[28]	0	1	2	0.13	3
10[29]	8	7	5	0.66	11
11[30]	4	2	16	1.5	6
12[23]	1	2	1	0.14	1
13[31]	13	0	39	0.59	33
14[30]	6	26	7	1.5	7
15[23]	0	5	6	0.033	6
16[32]	0	7	2	0.78	1
17[33]	8	5	14	0.58	14
18[33]	14	10	31	0.58	14
19[34]	0	0	2	0.086	0
20[35]	4	0	9	0.26	10
21[36]	8	45	36	0.38	38
22[37]	11	15	11	0.78	11
23[38]	0	0	0	0.062	0
24[39]	0	0	0	0.033	0
25[40]	0	0	1	0.031	0
26[41]	47	88	80	0.26	91
total:	**439**	**331**	**438**	**16.3**	**573**

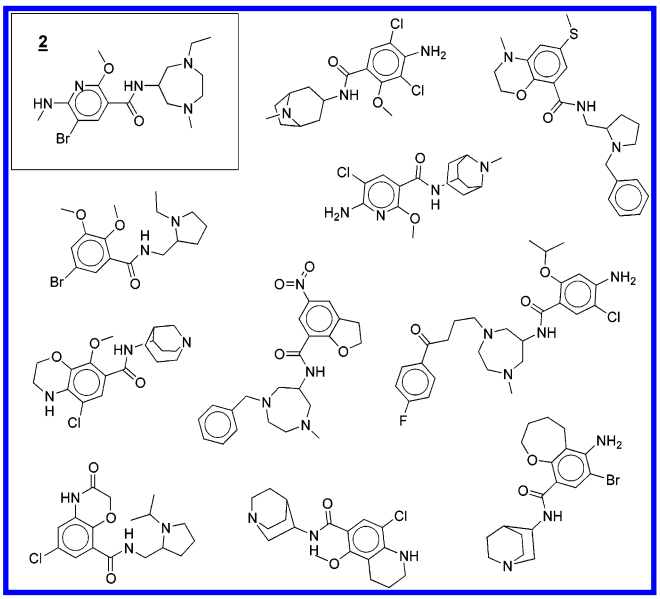

Figure 3. Example active structures retrieved for query *2*.

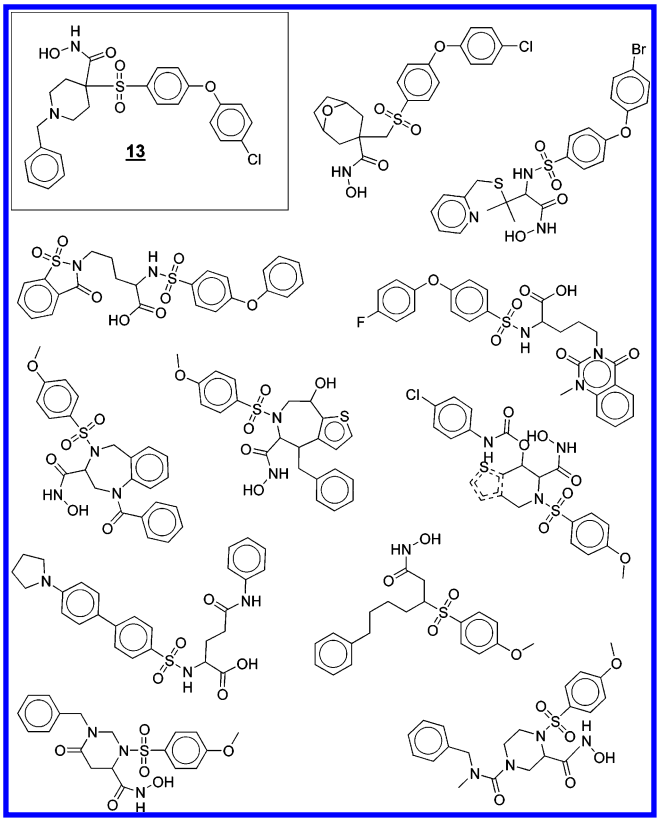

Figure 4. Example active structures retrieved for query *13*.

It is not possible to detail all of the various chemotypes retrieved by the CP screen here. However, Figures 3 and 4 depict some of the variety of active structures retrieved for queries *2* and *13*.

CONCLUSION

A new consensus scoring technique for fusing disparate virtual screening results called CP has been introduced. The proposed CP method was compared to the Sum-rank method previously evaluated in the literature where it was found to perform as well or better. It is also simpler to implement from a coding perspective. For a new virtual screening method to be considered, it must first be calibrated in order to generate its corresponding scoring function; however, this represents a small computational effort. The CP method is also limited by the fact that it considers each virtual screening score as an independent event from a probability perspective. A more appropriate approach would be to develop a CP scoring function based on joint probabilities. However, this is currently not possible due to practical considerations. It remains for future research whether heuristics can be developed to overcome this constraint.

The most appropriate combinations of screening methods for each fusion method were determined where a high degree of consistency between the CP and Sum-rank methods was observed. When fusing pairs of individual screening methods, it was found that combinations consisting of a fingerprint method (BCI or Daylight) and the 3D MCS performed the best for both the Sum-rank (equivalent assignment) and the proposed CP method. No statistically significant difference in effectiveness between the Sum-rank and CP method was observed when fusing pairs of screening methods.

When fusing three individual screening methods, it was discovered that four of the five best method combinations common to both fusion methods included one of the fingerprint methods (BCI or Daylight), one of the shape-based methods (ROCS or dbtop), and the 3D MCS. In addition, the other three method combination, BCI/2D MCS/3D MCS, also performed well. It was also found that the proposed CP method was superior to the Sum-rank method for three method combinations.

Although the proposed CP fusion method has been demonstrated using ligand-based virtual screening methods,

it is also applicable for fusing other screening methods such as docking scores and high-throughput screening (HTS) activity measurements provided enough reliable data is available for the probability calibration procedure. Unlike other fusion methods, the CP method is not dependent upon a monotonically increasing probability of activity with respect to increases in either a measured or calculated value.

ACKNOWLEDGMENT

The authors extend their gratitude to Barnard Chemical Information Ltd., Daylight Chemical Information Systems, OpenEye Software, and Tripos Inc. for their software support and to George Cowan and Kjell Johnson for helpful advice regarding the statistical analysis of the observed results. We also offer our appreciation to the reviewers for their helpful comments and suggestions.

REFERENCES AND NOTES

(1) Xu, H.; Agrafiotis, D. K. Retrospect and Prospect of Virtual Screening in Drug Discovery. *Curr. Topics Med. Chem.* **2002**, *2*, 1305−1320.
(2) Xue, L.; Bajorath, J. Molecular Descriptors in Cheminformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 363−372.
(3) Bajorath, J. Virtual Screening in Drug Discovery: Methods, Expectations and Reality. *Curr. Drug Discov.* **2002**, 24−28.
(4) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(5) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspect. Drug Discov.* **2000**, *20*, 1−16.
(6) Belkin, N. J.; Cool, C.; Croft, W. B.; Callan, R. K. Effect of Multiple Query Representations on Information System Performance. In *Proceedings of SIGIR*; Ed.; 1993; pp 339−346.
(7) Raymond, J.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.
(8) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.
(9) Whittle, M.; Willett, P. Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 449−457.
(10) Daylight Chemical Information Systems Inc. is at URL www.daylight.com.
(11) Barnard Chemical Information Systems Inc. is at URL www.bci.gb.com.
(12) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.
(13) Raymond, J.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Smoothed Bounded Distance Matrices. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 908−916.
(14) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer Similarity Searching of Conventional Structure Databases. *J. Mol. Graph. Model.* **2002**, *20*, 447−462.
(15) OpenEye is at URL www.eyesopen.com.
(16) Tripos Inc. is at URL www.tripos.com.
(17) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.
(18) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(19) Manmatha, R.; Rath, T.; Feng, F. Modeling Score Distributions for Combining the Outputs of Search Engines. In *Proceedings of the ACM SIGIR 01 Conference*; ACM, 2001; p 267−275.
(20) Mendenhall, W. Introduction to Probability and Statistics. Duxbury Press: 1987.
(21) Barreiro, E. J.; et al. Design, Synthesis, and Pharmacological Profile of Novel Fused Pyrazolo[4,3-*d*]pyridine and Pyrazolo[3,4-*b*][1,8]-naphthyridine Isosteres: A New Class of Potent and Selective Acetylcholinesterase Inhibitors. *J. Med. Chem.* **2003**, *46*, 1144−11152.
(22) Hirokawa, Y.; et al. Synthesis and Structure-Affinity Relationships of Novel N-(1-Ethyl-4-methylhexahydro-1,4-diazepin-6-yl)pyridine-3-carboxamides with Potent Serotonin 5-HT_3 and Dopamine D_2 Receptor Antagonistic Activity. *J. Med. Chem.* **2003**, *46*, 702−715.
(23) Silvestri, R.; et al. Simple, Potent, and Selective Pyrrole Inhibitors of Monoamine Oxidase Types A and B. *J. Med. Chem.* **2003**, *46*, 917−920.
(24) Lubisch, W.; et al. Benzoylalanine-Derived Ketoamides Carrying Vinylbenzyl Amino Residues: Discovery of Potent Water-Soluble Calpain Inhibitors with Oral Bioavailability. *J. Med. Chem.* **2003**, *46*, 2404−2412.
(25) Baraldi, P. G.; et al. Design, Synthesis, and Biological Evaluation of C^9- and C^2-Substituted Pyrazolo[4,3,-*e*]-1,2,4-triazolo[1,5-*c*]pyrimidines as New A_{2A} and A_3 Adenosine Receptors Antagonists. *J. Med. Chem.* **2003**, *46*, 1229−1241.
(26) Aranapakam, V.; et al. Synthesis and Structure−Activity Relationship of α-Sulfonylhydroxamic Acids as Novel, Orally Active Matrix Metalloproteinase Inhibitors for the Treatment of Osteoarthritis. *J. Med. Chem.* **2003**, *46*, 2361−2375.
(27) Jansen, M.; Potschka, H.; Brandt, C.; Loscher, W.; Dannhardt, G. Hydantoin-Substituted 4,6-Dichloroindole-2-carboxylic Acids as Ligands with High Affinity for the Glycine Binding Site of the NMDA Receptor. *J. Med. Chem.* **2003**, *46*, 64−73.
(28) Mitsumori, S.; et al. Synthesis and Biological Activity of Various Derivatives of a Novel Class of Potent, Selective, and Orally Active Prostaglandin D2 Receptor Antagonists. 1. Bicyclo[2.2.1]heptane Derivatives. *J. Med. Chem.* **2003**, *46*, 2436−2445.
(29) Hah, J. M.; Martasek, P.; Roman, L. J.; Silverman, R. B. Aromatic Reduced Amide Bond Peptidomimetics as Selective Inhibitors of Neuronal Nitric Oxide Synthase. *J. Med. Chem.* **2003**, *46*, 1661−1669.
(30) Mylari, B. L.; et al. A Highly Selective, Non-Hydantoin, Non-Carboxylic Acid Inhibitor of Aldose Reductase with Potent Oral Activity in Diabetic Rat Models: 6-(5-Chloro-3-methylbenzofuran-2-sulfonyl)-2*H*-pyridazin-3-one. *J. Med. Chem.* **2003**, *46*, 2283−2286.
(31) Aranapakam, V.; et al. Synthesis and Structure−Activity Relationship of N-Substituted 4-Arylsulfonylpiperidine-4-hydroxamic Acids as Novel, Orally Active Matrix Metalloproteinase Inhibitors for the Treament of Osteoarthritis. *J. Med. Chem.* **2003**, *46*, 2376−2396.
(32) Gitto, R.; et al. Discovery of a Novel and Highly Potent Noncompetitive AMPA Receptor Antagonist. *J. Med. Chem.* **2003**, *46*, 197−200.
(33) Choi-Sledeski, Y. M.; et al. Discovery of an Orally Efficacious Inhibitor of Coagulation Factor Xa Which Incorporates a Neutral P1 Ligand. *J. Med. Chem.* **2003**, *46*, 681−684.
(34) Wood, M. R.; et al. Benzodiazepines as Potent and Selective Bradykinin B1 Antagonists. *J. Med. Chem.* **2003**, *46*, 1803−1806.
(35) Wyss, P. C.; et al. Novel Dihydrofolate Reductase Inhibitors. Structure-Based Versus Diversity-Based Library Design and High-Throughput Synthesis and Screening. *J. Med. Chem.* **2003**, *46*, 2304−2312.
(36) Schmidt, J. M.; et al. De Novo Design, Synthesis, and Evaluation of Novel Nonsteroidal Phenanthrene Ligands for the Estrogen Receptor. *J. Med. Chem.* **2003**, *46*, 1408−1418.
(37) Silvestri, R.; et al. Novel Indoyl Aryl Sulfones Active Against HIV-1 Carrying NNRTI Resistance Mutations: Synthesis and SAR Studies. *J. Med. Chem.* **2003**, *46*, 2482−2493.
(38) Bohme, T. M.; et al. Structure−Activity Relationships of Dimethindene Derivatives as New M_2-Selective Muscarinic Receptor Antagonisits. *J. Med. Chem.* **2003**, *46*, 856−867.
(39) Zhuang, L.; et al. Design and Synthesis of 8-Hydroxy-[1,6]Naphthyridines as Novel Inhibitors of HIV-1 Integrase in Vitro and in Infected Cells. *J. Med. Chem.* **2003**, *46*, 453−456.
(40) Yu, G.; et al. Substituted Pyrazolopyridopyridazines as Orally Bioavailable Potent and Selective PDE5 Inhibitors: Potential Agents for Treatment of Erectile Dysfunction. *J. Med. Chem.* **2003**, 457−460.
(41) Beukers, M. W.; et al. N^6-Cyclopentyl-2-(3-phenylaminocarbonyltriazene-1-yl)adenosine (TCPA), a Very Selective Agonist with High Affinity for the Human Adensoine A_1 Receptor. *J. Med. Chem.* **2003**, *46*, 1492−1503.

CI034234O